

F

Facades and Enclosures, Building for Sustainability

VOLKER HARTKOPF¹, AZIZAN AZIZ¹, VIVIAN LOFTNESS²

¹Center for Building Performance and Diagnostics, Carnegie Mellon University, Pittsburgh, USA

²Center for Building Performance and Diagnostics, Carnegie Mellon University, Pittsburgh, PA, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Facades that Maximize Individual Access to the Natural Environment

Facades that Ensure Daylighting as the Dominant Source for Both Task and Ambient Lighting

Facades that Ensure Natural Ventilation as the Dominant Fresh Air and Cooling Source Combined with Mixed-Mode HVAC Conditioning

Facades that Minimize Enclosure Heat Loss/Heat Gain

Facades that Integrate Climate-Responsive Shading and Glare Control

Facades that Engineer Load Balancing and Mean Radiant Temperature Control

Facades that Integrate Passive and Active Solar Heating, Cooling, and Power

Facades Designed to Manage Water

Facades that Maximize Enclosure Life

Future Directions: Facade Innovation Through Systems Integration

Bibliography

Glossary

Facade enclosure envelope Throughout this entry, Facade is a subset of Enclosure and Enclosure is considered the same as Envelope.

Carbon neutral net-zero and the 2030 challenge

Carbon-neutral buildings and net-zero buildings are typically defined as buildings designed to achieve net-zero in green house gas emissions during their lifetime. This has been set as a requirement for all new buildings by 2030 in the 2030 Challenge™ (url).

Passive survivability Environmental Building News uses the term passive survivability to describe a building's ability to maintain critical life-support conditions in the event of extended loss of power, heating fuel, or water, or in the event of extraordinary heat spells.

Biophilia Introduced by E.O. Wilson, the biophilia hypothesis suggests that there is an instinctive bond between human beings and other living systems that must be met by buildings that ensure critical connections.

Ambient and task lighting Ambient lighting is the light needed for safe movement and non-visual tasks, while task lighting is linked to visual acuity at task and must be glare-free. Daylight can be used for both, but requires significantly greater design resolution for task lighting.

Mixed-mode or Hybrid HVAC An approach to space conditioning that combines natural ventilation from operable windows or vents with mechanical heating, cooling, and ventilation systems (HVAC).

Night ventilation An approach to space conditioning that utilizes cool night air to precool the buildings' thermal mass in order to reduce daytime cooling demands.

Thermal bridge A conductive or non-insulated element in the facade assembly that allows heat to be unnecessarily lost or gained, and allows condensation to occur, to be addressed by well-designed thermal breaks.

PassivHaus™ A standard that codifies energy conservation and passive conditioning requirements for residential and small commercial buildings.

Double envelope The introduction of two facades with a captured air space and static or dynamic components that support orientation- and season-specific modulation of climate conditions.

Rainscreen An independent layer on the facade to manage weather challenges, separated by vent spaces from the structural, thermal, and vapor management of the primary facade.

Definition of the Subject

Sustainable buildings have enclosures that ensure the highest level of thermal, visual, acoustic, and air quality, as well as design for change, through system flexibility, and for long-term building integrity (Fig. 1). Sustainable enclosure design clearly demonstrates advances in enclosure components and assembly design, as well as their effective integration with interior systems, structural systems, mechanical conditioning systems, lighting systems, telecommunications and power systems. Sustainable enclosures effectively address ten major responsibilities, indeed areas of major innovative opportunity, in the pursuit of aesthetics, comfort, and resource conservation:

- 1. Access to the natural environment
- 2. Daylighting

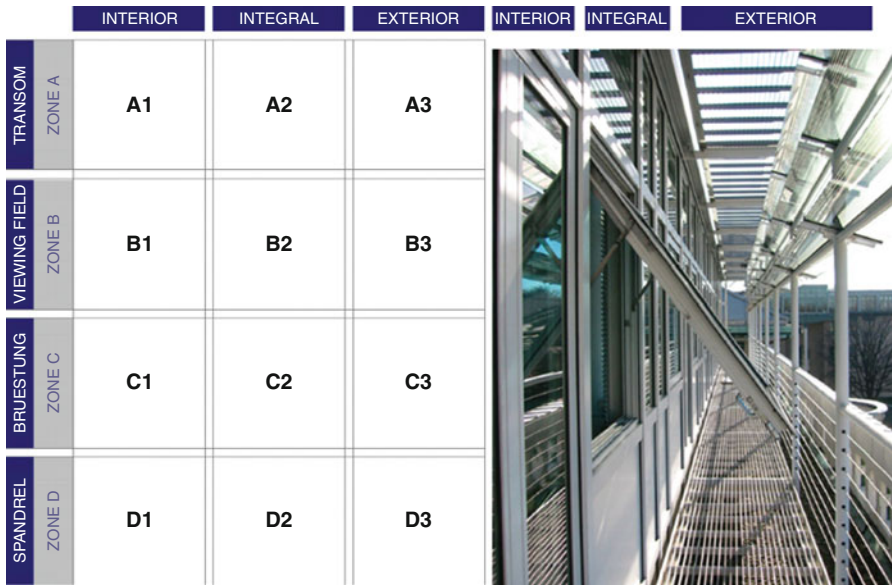
- 3. Natural ventilation
- 4. Heat loss/heat gain control
- 5. Solar control
- 6. Load balancing
- 7. Passive and active solar
- 8. Water
- 9. Enclosure life
- 10. Systems integration

In addition to performance goals for building roofs and exposed floors, the design of high-performance building enclosures must fully resolve at least 12 facade layers: the transom, viewing field, kick plate, and spandrel (top to bottom) and the interior, integral and exterior layers (inside to out) to resolve each of the 10 performance goals (The relative importance of facade components and layers to each performance outcome will be highlighted in graphics with deepest shading for the most important design decisions.)

Introduction

Architecture Unplugged: Superefficient Buildings that Environmentally Surf

Carbon-neutral buildings, net-zero buildings, and the 2030 challenge. To meet these goals, the next generation



Facades and Enclosures, Building for Sustainability. Figure 1
Diagram illustrating 12 areas in the facade critical to various performance outcomes

of buildings must be designed to use so little energy that on-site renewable energy can fully meet the heating, cooling, and lighting loads – loads that represent over 35% of all US energy.

For the first generation of projects, these challenges are being met by tight buildings, with superefficient equipment, fed in part by solar installations. While this will help ensure a 30% reduction, it will never achieve carbon neutrality.

Carbon neutrality is critically dependent on the maximum use of nature's renewable energies – daylight, natural ventilation, natural cooling, and passive solar heating. Building systems will need to be turned off as long as possible – ensuring buildings that “free-roll” or better yet “environmentally surf” through hours, days, months, and seasons with natural conditioning. The beauty of free-rolling buildings goes beyond energy and carbon, however, to support health, productivity, and a higher quality of life.

The most critical early design decisions relate to building massing – the height, depth, and orientation of buildings – which has a major impact on indoor environmental quality and energy loads. Neither “the tallest building in the world” nor “the largest building under one roof” will offer sustainable work environments in a power failure. Indeed, these buildings create significantly higher-energy loads in almost every climate since they eliminate the use of daylight, natural ventilation, or the natural dissipation of internal heat gains through the building skin [1], and must be abandoned in a power outage. If air quality, comfort, and energy are drivers for building form, then the next generation of buildings will strive toward controlling the building depth, height, and orientation to achieve environmental comfort for a maximum percentage of the year – unplugged.

“Architecture Unplugged” would require enclosures that demonstrate serious attention to the management of solar gain, heat transfer, moisture migration, and day-night load balancing – a form of environmental surfing that depends on dynamic facades that recognize time of day, season, and variability of climate. These mass, color, venting, and thermal insulation characteristics are key to energy conservation in buildings, and require entirely regional solutions. Heavy capacitance masonry facades will be seen in desert climates with large diurnal swings; heavily

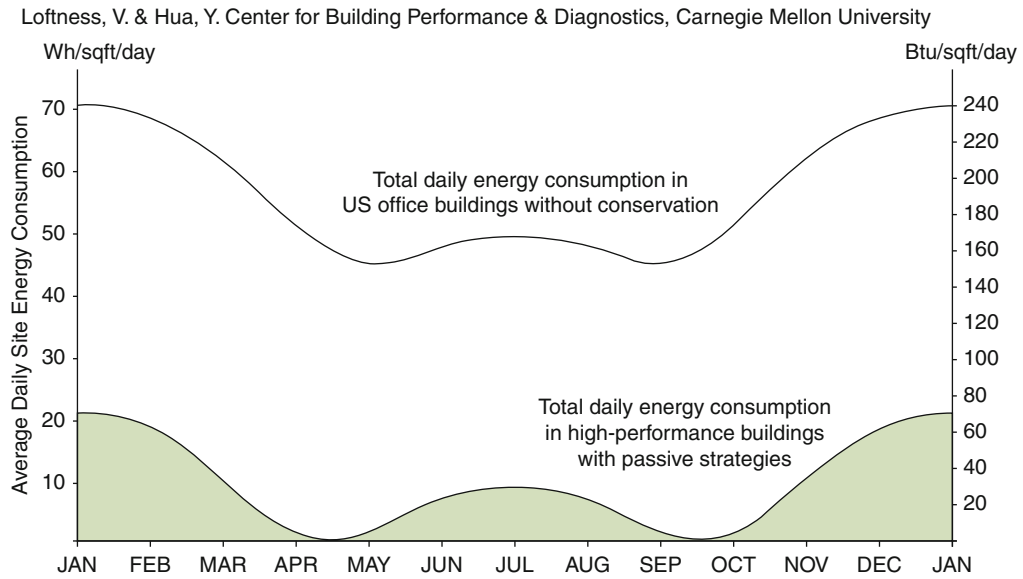
shaded and highly operable facades will be seen in warm, humid climates; and solar exposed yet well-insulated facades will be seen in cold climates. In each climate, the massing, orientation, and selection of facade materials will be regional in character, providing the maximum natural comfort before mechanical systems need to be introduced (Figs. 2 and 3).

Facades that Maximize Individual Access to the Natural Environment

The most significant step that designers, engineers, and building managers can take to move toward healthier and more sustainable buildings is the commitment to increasing individual contact with the outside environment. The emerging field of “biophilia” has identified the importance of access to nature to human well-being [2–5], exploring a range of qualities of the natural environment: views, daylight, sunlight, fresh air, breezes, access to outdoor spaces and activities, circadian rhythms, seasonal and daily climate variations, and nature's sounds, smells and habitat. Windows and doors connect building occupants with a richness that may be critical to individual health. At the same time, windows and doors provide those outside the building with a level of transparency, oversight, and contact that may be critical to community.

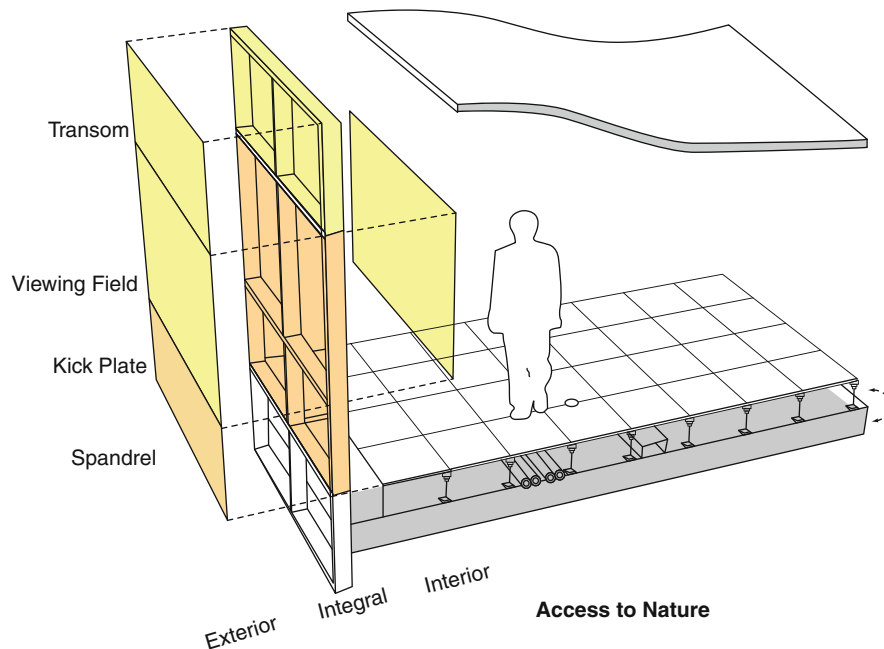
Europe and Scandinavia have guaranteed every worker seated access to a window with views, setting a maximum distance of 7 m, or 20 ft from the window wall. In Switzerland and Scandinavia, they have further guaranteed access to *operable* windows in each workplace [6]. With these standards, buildings typically do not exceed 60 ft from window wall to window wall, and always maintain operable, dynamic facades instead of sealed, seasonally unchanging building enclosures.

Driven by the advent of air conditioning and inexpensive electricity, the “obese” buildings of today (maximum volume – minimum surface area buildings) have actually resulted in a dramatic increase in the length of the cooling “season” for buildings in all climates. They have also resulted in a significant increase in lighting loads during daylight hours, in combination often contributing more than 50% of the total energy demand in buildings. The study of the Building



Facades and Enclosures, Building for Sustainability. Figure 2

Deep energy conservation combined with passive conditioning including daylighting, natural ventilation, and passive solar can dramatically reduce energy consumption



Facades and Enclosures, Building for Sustainability. Figure 3

Critical facade elements to maximize access to the natural environment

Research Establishment of a portfolio of UK office buildings clearly illustrates the energy penalty of sealed, deep floor plan “prestige” buildings [7].

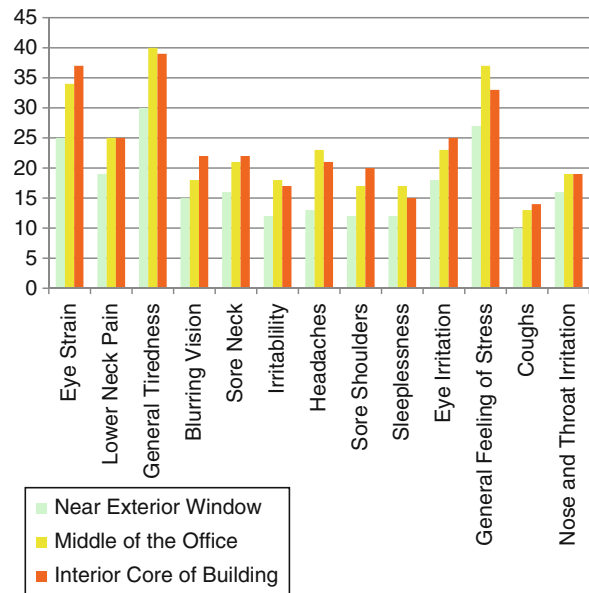
While there is significant debate about the importance of indoor daylight or sunshine for human health or performance, there seems to be a growing consensus that access to views is significant. Beginning with the seminal work of Ulrich [8], then Mendell [9], Heschong Mahone Group [10], and now Kellert [4], seated views of nature and proximity to windows are being linked to reduced length of stay after surgery, reduced sick building syndrome, increased performance at task, and overall improved emotional health.

In a 1990 survey of over 2,000 employees in two buildings at the US Department of Energy, Carnegie Mellon University’s Center for Building Performance identified 10–20% lower sick building symptoms among employees with seated views of windows, controlling for rank [11]. Whether user perception of personal health is improved due to the light, the view, the perimeter conditioning systems, or an increased level of environmental control (blinds, HVAC controls) at the window is unclear. Regardless, there is a measurable benefit to ensuring that a workforce has fewer health symptoms across the board (Fig. 4).

Two case studies that frame the conclusion that views are a significant factor in health and productivity are captured below from the Carnegie Mellon Building Investment Decision Support Tool (BIDSTTM) [7, 8, 10]:

Window View of Nature = Health

In a 1984 observational field study at a Pennsylvania hospital, Ulrich identifies an 8.5% reduction in postoperative hospital stay (8.7 days vs. 7.96 days) for gall bladder surgery patients who had a view of a natural scene from their hospital room, as compared to those with a view of a brick wall. Patients with a view of a natural scene also received fewer negative evaluations from nurses and took fewer strong analgesics (Pennsylvania Hospital/Ulrich [8]).



Facades and Enclosures, Building for Sustainability.

Figure 4

Window proximity correlates to a 5–25% reduction in health complaints among 2,000 workers in two US office buildings, controlled for rank [11]

Window View of Nature = Productivity

In a 2003 building case study of the Sacramento Municipal Utility District (SMUD) Call Center, Heschong et al. identify a 6–7% faster Average Handling Time (AHT) for employees with seated access to views through larger windows with vegetation content from their cubicles, as compared to employees with no view of the outdoors (SMUD/Heschong Mahone Group [10]).

These studies illustrate the value of windows with views for all occupied spaces, alongside effective design of window size and orientation, sight lines, and view content. The studies suggest that views of the outdoors should be with “biophilic” content – views of pedestrians and trees and community life – to critically maintain our sense of time and season. At the same time, it will be imperative for designers to ensure that requirements for glare control, heat gain and noise

control, privacy and security are equally met—the definition of quality design (Fig. 5).

In pursuing design for sustainability, the challenge to designers is not only to connect indoor spaces with the outdoors visually, but to physically connect building occupants with nature – each site’s unique climate and seasons, the textures, sounds, smells, and diversity of landscape and species. This not only suggests a shift away from mega-plexes and high-rise buildings toward open-air campus and village planning, but a commitment to operable windows, and distributed doors, terraces, and landscapes.

The physiological and psychological benefits of an ongoing connection to the locally unique and seasonally dynamic natural environment still need quantification. For example, studies in early childhood development reveal the academic and emotional benefits of natural playgrounds as opposed to hard-top play yards [12]. Sustainable buildings must incorporate physical access to outdoor spaces for a host of reasons, including physical exercise, exposure to sunshine and circadian rhythms, variations in thermal stimuli, and access to community (Fig. 6).

Facades that Ensure Daylighting as the Dominant Source for Both Task and Ambient Lighting

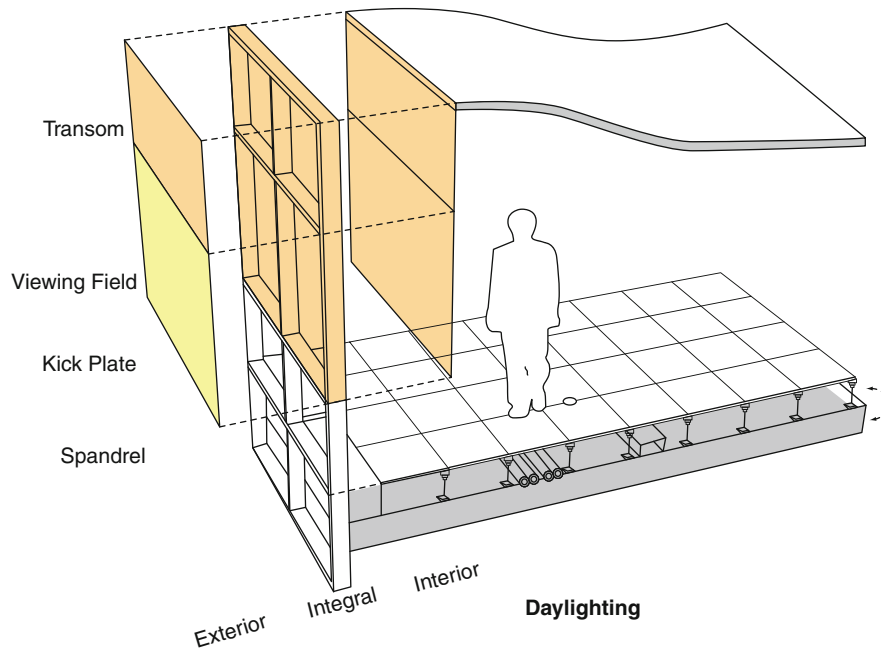
- “A building with lights on during the daytime is not a sustainable building.”

Electric lighting is almost 25% of our primary energy use and over 35% of all electric energy use in the USA [13] – and much of that is during the daytime! Energy-efficient lamps, ballasts, and fixtures are obvious first steps for ensuring 30% lighting energy savings, and daylight responsive and occupancy responsive controls for these fixtures can reduce the next 20% (reference: see [14]). The giant leap, however, is achieved by buildings designed for daylighting as the *dominant* light source—environmentally “surfing” for 90% of the day without electric demand for lighting. This is more than rediscovering the design expertise of the nineteenth century. Good daylighting design demands effective integration of new glazing technologies, introducing light redirection and shading layers that enrich and regionalize architecture. Critically, buildings must be articulated, in finger or courtyard shapes, to ensure universal access to daylight.



Facades and Enclosures, Building for Sustainability. Figure 5

Designing window dimensions, orientations, and sight lines as well as adjacent landscaped spaces (with substantial storm water management benefits) are critical to ensuring access to views with “biophilic” content



Facades and Enclosures, Building for Sustainability. Figure 6

Critical facade elements to ensure daylighting as the dominant light source

There are two main alternatives in the use of daylight in work environments. The first is the use of daylight for ambient lighting requirements of 200–300 lx and for spatial highlights. The second is the use of daylight as the dominant task lighting source at the work surface. This second alternative offers the greatest energy savings but also the greatest design challenge, since both the daylight “fixture” and the electric lighting interface need to be effectively designed. The daylight fixture – window, transom, or skylight – will need appropriate orientation, size, reflector-frame design and lens-blind design as well as the corresponding design of the room to ensure effective light distribution without glare.

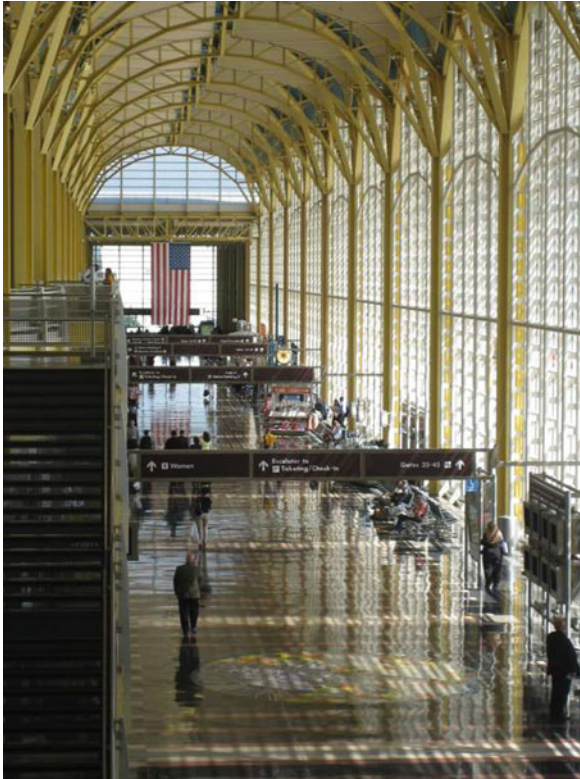
There are a broad range of traditional and innovative technologies to be considered for effective daylighting of buildings to achieve both ambient and task lighting needs. New developments in glazing enable architects to specify high visual transmittance (short wave energy) while controlling solar transmittance and heat loss (long wave energy). Given innovations in glazing technology, it is possible to achieve high light transmittance over 50% while maintaining shading coefficients below 40%

in cooling load–dominant climates, or above 60% in heating load–dominant climates.

However, sustainable design recognizes that not all shading should be solved with the glass selection, since this can rob spaces of daylight and solar heat where needed.

Each window must be designed as a lighting “fixture” in relation to space configuration and layout. If lighting engineers can design effective reflectors and diffusers for point and linear sources, they can design for planar sources as well. Transom glazing will be critical to effective distribution, combined with exterior, integral, or interior light redirection devices to ensure a more even distribution of daylight and to reduce glare from the window. If there is a shortage of window access, it should never be given to circulation aisles in a sustainable design (Fig. 7).

The rules for effective daylighting in offices, schools, and hospitals have long been established, relating to percent of aperture, ceiling height, room depth and color, as well as glare control and light redirection devices at the window. Today, high-performance



Facades and Enclosures, Building for Sustainability.

Figure 7

Daylight is celebrated at Reagan National Airport, Washington DC

enclosures go even further, fully exploring the benefits of a layered facade. The design of layered facades enables seasonally and daily dynamic control of the light, heat, and ventilation energies available in the natural environment. All three layers should be designed together – external shading and light shelves to distribute daylight and reduce high sun angle glare; integral high visible transmittance glazing with climate responsive or dynamic shading coefficients; and internal light redirection blinds that sustain views while diminishing low sun angle glare. It is critical that the design of these layered facades be entirely regional in character, and suited to the function of the building, as they control light, glare, solar heat, heat loss/gain, even thermal mass. By displacing mechanical and electrical loads, these facades provide long-term energy savings. In retrofitting deep buildings for daylighting,

the sustainable design community is exploring atria, light wells, light ducts, and light pipes. Bank of America in San Francisco cut an atrium into an overly deep building to bring daylight to all spaces. Innovations in light pipes, reflective ducts, and tracking/beaming heliostats all have potential for increasing daylight effectiveness and delight (Fig. 8).

At the same time, the daylighting “system” must give the appropriate information to the electric lighting interface so that it can offer the appropriate “infill” through continuous dimming control. Studies indicate that the selection, placement, and installation of the photo-sensor is critical, concluding that a ceiling-mounted photosensor shielded from direct window light will ensure the best correlation with daylight work-plane illuminance [15]. The net impact of daylighting and electric lighting integration is significant energy savings [16] and improved workplace satisfaction.

Emerging research is revealing that day-lit offices, classrooms, and hospitals may measurably contribute to greater health and performance at task. Light levels can be higher without energy penalty; full spectrum light offers rich color rendition and 3-dimensional modeling; circadian rhythms set by daylight variations throughout the day trigger melatonin production and sleep patterns; and views meet fundamental human needs.

Sustainable buildings are designed for daylight to meet all ambient and task requirements during the day. The investment in a layered facade for daylight effectiveness will simultaneously support natural ventilation, night ventilation, thermal load balancing, and passive solar heating – the full complement of regional opportunities for natural conditioning.

Daylight Lighting Control = Energy Savings

In a 1997 before and after building case study of the New York City federal building, the Electric Power Research Institute (EPRI) identifies 64% lighting energy savings and net HVAC energy savings of 0.07 kWh per square foot annually following the installation of an energy-efficient lighting system with daylight dimming and lumen maintenance controls (NYC Federal Building/ EPRI [16]).



Facades and Enclosures, Building for Sustainability. Figure 8

The central atrium and “sky gardens” at Commerzbank in Frankfurt bring daylight to all spaces and provide occupants with access to nature in a dense urban environment

Daylight Spectrum and Timing = Individual Productivity

In a 1997 controlled experiment, Boyce et al. identify a 1.6–12.8% improvement in night-shift workers’ performance on short-term memory and logical reasoning tasks under large skylight-simulating fixtures with hidden fluorescent lamps, capable of providing fixed or variable illuminance from 200 lx to 2800 lx. Performance was enhanced by fixed high illuminance of 2800 lx and by steadily decreasing illuminance that simulated daylight from midday to dusk, as compared to fixed low illuminance of 250 lx or steadily increasing illuminance that simulated daylight from dawn to midday (Boyce et al. [17]).

Facades that Ensure Natural Ventilation as the Dominant Fresh Air and Cooling Source Combined with Mixed-Mode HVAC Conditioning

► “A sealed building is not a sustainable building.”

The use of *natural ventilation and natural cooling* with direct outside air for swing and winter seasons ensure affordable high levels of fresh air with the greatest energy efficiency (Fig. 9). International studies consistently link increases in outdoor air supply to both productivity and health gains in the workplace [9].

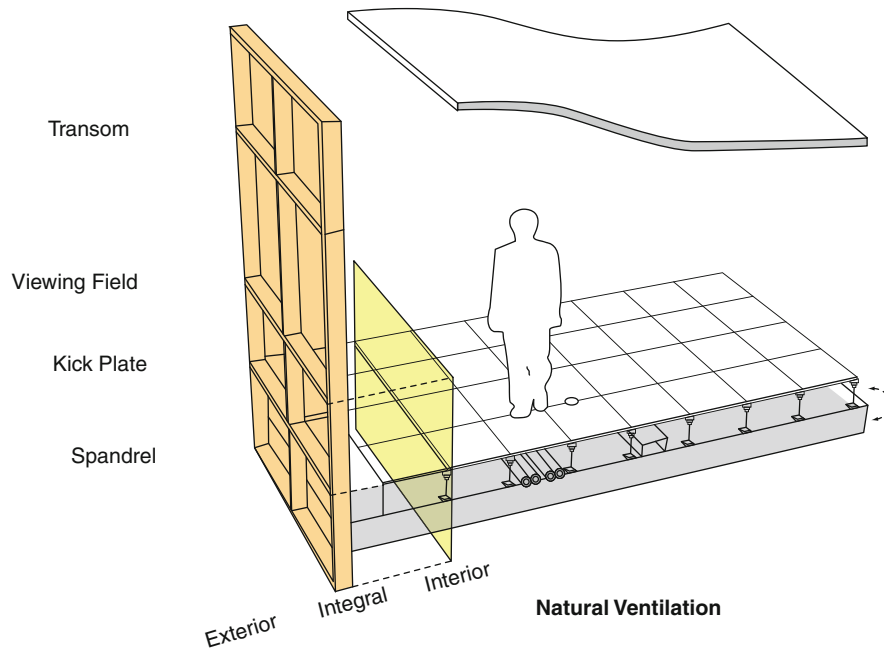
Natural Ventilation = Health

In a 1991 multiple building study of 880 workers in eleven office buildings in the San Francisco bay area, Mendell identified 20–81% reductions in reported SBS symptoms for occupants of naturally ventilated buildings as compared to occupants of office buildings with air conditioning or mechanical ventilation (CA Healthy Building Study/ Mendell [9]).

Natural Ventilation = Health + Individual Productivity

In a 2004 multiple building study of professional middle-aged women in France, Preziosi et al. identify a 57.1% reduction in absenteeism, a 16.7% reduction in medical services use (doctor visits), and a 34.8% reduction in hospital stays among subjects with natural ventilation in their workplace, as compared to those with air conditioning (Preziosi et al. [18]).

In addition, the substantial opportunity for “free cooling” through the use of outside air is significantly underestimated by calculations focused on only “comfort zone” natural ventilation. Natural cooling is possible whenever outside conditions are within or below the comfort zone both in the day and at night, as long



Facades and Enclosures, Building for Sustainability. Figure 9

Critical facade elements to ensure natural ventilation as the dominant fresh air and cooling source

as drafts can be effectively managed through mixing or heat recovery. The use of both passive (open windows) and active economizer cycles can save from 50% to 100% of the cooling demand in various regions of the USA, when including the benefits of nighttime cooling of the building's thermal mass [19].

Natural Ventilation = Individual Productivity + Energy Savings

In a 2002 multiple building study of 39 Australian office settings, Rowe identifies a 13.8% perceived productivity improvement, a 15% improvement in overall comfort, and 79% annual HVAC energy savings due to replacing a mechanical air-conditioning system with mixed-mode HVAC that supports both natural ventilation and air conditioning (Australian Offices/ Rowe [20]).

Design for effective natural ventilation and natural cooling of the building mass and its occupants requires multi-disciplinary expertise to ensure that building shape, orientation, atria/courtyards, chimneys, window

types (awning, casement, double hung), sizes, and locations ensure effective cross and stack ventilation. Cross and stack ventilation demands that building form and space layout reflect prevailing wind directions and the potential for thermal and solar assist, often studied through wind model testing or computational fluid dynamic simulations. In climates with day-night temperature swings, night ventilation should be explored and building materials or storage selected to store cooler temperature, creating a thermal “flywheel” for the following day.

In many climates, “mixed-mode” HVAC systems are designed to support natural ventilation and to provide air conditioning and ventilation – when and where natural ventilation is inadequate. There are three types of mixed-mode HVAC systems [7]:

- Concurrent systems, which use natural ventilation and mechanical HVAC simultaneously. Occupants are free to open windows and the HVAC system provides supplemental ventilation, dehumidification, and cooling, while an advanced control system coordinates zone air supply rates with window positions.

- Changeover systems, where the building alternates between natural and mechanical mode on a seasonal or daily basis.
- Zoned systems, in which different conditioning strategies are used simultaneously in different zones of a building.

Concerns related to outdoor air pollution, pollen, and humidity can be effectively addressed with well-designed mixed-mode conditioning systems (Fig. 10). Natural ventilation in high-rise buildings demands both architectural and engineering innovations to manage both pressurization issues and stack effect. This has been accomplished through the design of stacked 6–8-story buildings such as the Commerzbank in Frankfurt, and through double envelope designs, where wind and pressure are moderated through intermediate facade spaces.

For both natural ventilation and mechanically assisted ventilation, it is critical to engineer *ventilation effectiveness*, ensuring adequate fresh air volumes through large apertures or economizers; accessible,

large volume fresh air paths from air-intake locations to occupants, such as in displacement ventilation systems; multiple fresh air sources with operable windows, distributed air handlers or facade integrated HVAC; and adequate exhaust without recirculation or contamination of the air flow paths. Air intakes especially should be strategically located, not near loading docks with idling engines or on roof decks with standing water, but in non-polluted areas, without standing water or solar overheating. Indeed, air sources should be freshest where prevailing breezes travel through shaded, landscaped areas with some level of natural air purification and cooling.

To achieve effective natural ventilation with mixed-mode mechanical conditioning, the design team must work together to engineer and integrate the natural ventilation/natural cooling controls and the HVAC controls to maximize air quality, user control, passive conditioning, and energy efficiency. Given a growing body of research illustrated below, natural ventilation with mixed-mode conditioning should be the norm in



Facades and Enclosures, Building for Sustainability. Figure 10

Operable windows at Commerzbank, a high-rise structure, provide natural conditioning for the occupants, linked with mechanical cooling in a hybrid system. The atrium at Institute for Forestry and Nature Research in Wageningen provides a protected space for natural ventilation in the offices while giving occupants a natural setting for meetings and relaxation

all commercial buildings for human health and productivity, energy efficiency, and the viability of occupied spaces in the face of electricity shortages.

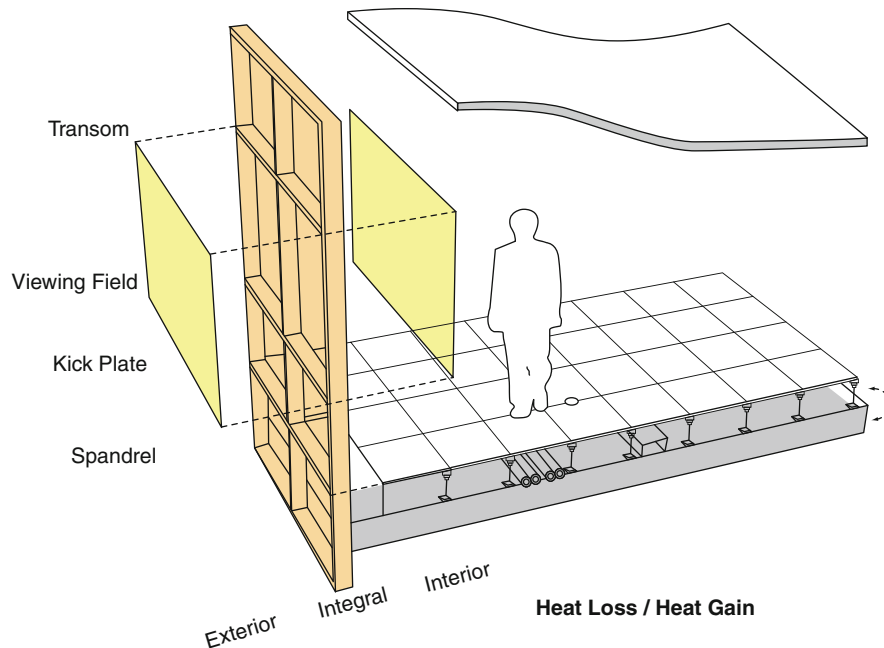
Facades that Minimize Enclosure Heat Loss/Heat Gain

One of the earliest decisions affecting the sustainability of buildings is the overall massing of the building and the *orientation of major facade and roof surfaces* (Fig. 11). While very harsh climates might benefit from minimum surface area to usable space enclosed, there are climate-specific arguments for increasing building enclosure for daylighting, natural ventilation, passive solar heating, and even the dissipation of excessive internal heat gains (see [Load Balancing](#)). These increased exposures must be strategic, however, avoiding solar overheating, cold air infiltration, and increased heat loss or heat gain. Serious re-evaluation of vast roof exposures and cantilevered floors is needed, minimizing excessive heat losses or heat gains unless they have corresponding benefits from natural conditioning. Even the location of the HVAC system will be significant, with rooftop and west facing cooling

equipment measurably less efficient due to solar heat gains. Design goals should be set for building shape and orientation to ensure the best integration of passive conditioning while reducing environmental loads.

Architects make major decisions about the thermal performance of buildings through the specification of *window to wall ratios*. Sustainable designers weigh heat loss and heat gain against daylighting, natural ventilation, passive solar heating, and load balancing through the facade, alongside the spatial flexibility needed by dynamic organizations. Innovative building enclosures are designed to act like the human skin, dynamically supporting heat, air, and moisture dissipation for the health of the occupants in changing climates. The sustainable design community has rediscovered the opportunities for dynamic thermal performance of building enclosures, with high-resistance louvers, curtains, and dynamic layered facades changing heat transfer characteristics hour by hour, day to night, or season by season.

Needless to say, *the highest level of thermal resistance* to heat loss and heat gain is pursued for both the wall and window assemblies in sustainable buildings. Innovations in wall insulation have supported shifts from



Facades and Enclosures, Building for Sustainability. Figure 11
Critical facade elements to minimize enclosure heat loss/heat gain

R5 per inch to R10 per inch – a doubling of performance specifications. Window innovations also abound, with R2 double glazed assemblies replaced by R4, R8, and even R12 assemblies as a result of low-e, gas filled assemblies with innovative spacer and frame technologies, even while visible transmission is critically kept above 50%. At the same time, low solar heat gain coefficients (SHGC) of 0.35–0.50 can be achieved where needed to reduce overheating at the facade without loss of visual connection to the outside. At a very minimum, these three window variables – resistance to heat transfer (R), visible transmission (t_v), and solar heat gain coefficients (SHGC, solar transmission) – must be balanced in response to each climate and building type.

Facade Temperature Control = Individual Productivity

In a 2000 field case study of telecommunication office workers in Finland, Hannula et al. identify a 2.8% increase in measured productivity in north-facing offices with an average temperature of 23.6°C (range 21.9–27.8°C) as compared to south-facing offices with an average temperature of 25.2°C (range 22.8–28°C), supporting the need for improved heat gain control by orientation (Telecommunication Office/ Hannula et al. [21]).

Finally, sustainable enclosure design detailing ensures the *lowest level of thermal bridging* through walls, roofs, floors, and foundations. Structural and enclosure designers must work together to ensure the elimination or minimization of structural elements that connect indoor heated or cooled areas with the outdoors, including the facade, roof, floor, and foundation. Ideally the thermal “wrapper” is continuous and fully outside the structural frame, except at foundation connections that might indeed support ground source conditioning (see [Thermal Load Balancing](#)). Cantilevered balconies, overhangs, floors, and roof elements can contribute to significant heat loss, moisture and vapor migration, and condensation that lead to building degradation (see [Enclosure Life](#)). Not only the structural integration, but also product specification, enclosure detailing, and design for constructability are

critical to the thermal performance of facades. There are numerous conditions where well-insulated facades, roofs, and foundations are compromised by penetrations that were not anticipated in the drawings (vents, tie rods, utility meters, etc.), or not adequately sequenced in construction to ensure the continuity of the thermal envelope.

One of the most exciting new developments in enclosure design is *green roofs and green walls*. Green roofs and their layers of waterproofing, insulation, soils, and vegetation have the potential to improve roof resistance to heat gain by 50% (NRC reference – see [22]), improve roof longevity by a factor of 2 [23], reduce storm water runoff by 50% [24], reduce air plane noise transmission, as well as reduce urban heat islands and air pollution [25, 26]. Green walls are also offering measurable gains by reducing solar heat gain, providing evaporative cooling, and creating seasonal qualities for our buildings (Fig. 12).

Green Walls + Double Skin Façade = Energy Savings

In a 2005 controlled lab experiment in the Netherlands, Stec et al. identify an 18% reduction in cooling system capacity requirements and a 19% reduction in cooling energy consumption for a double skin façade with plants in the cavity rather than blinds, due to a 20% reduction in temperature on the interior wall of the façade (Stec et al. [27]).

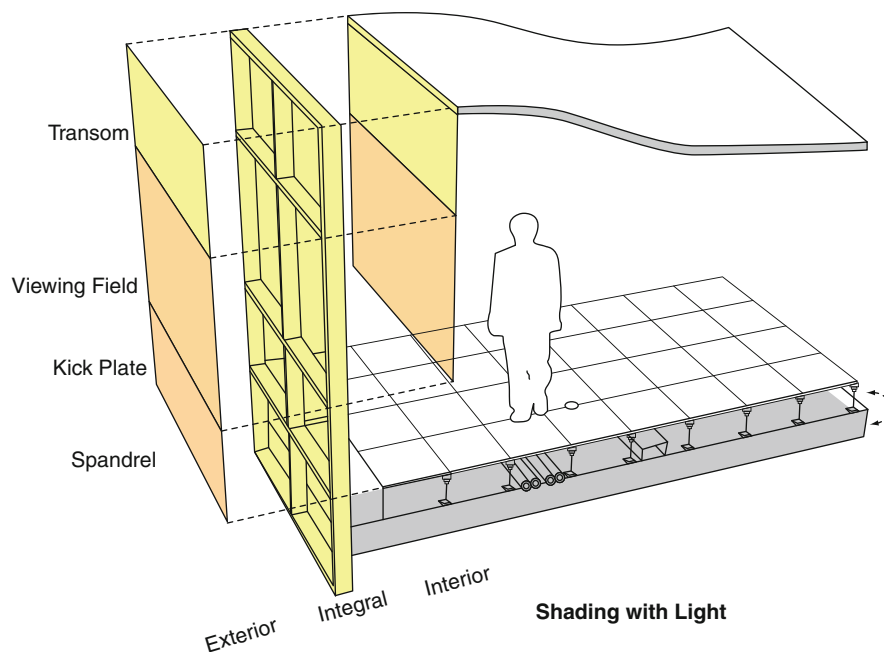
Facades that Integrate Climate-Responsive Shading and Glare Control

As with heat loss and heat gain, managing solar overheating and glare begins with the careful *sizing and location of windows* to ensure shade and light without glare (Fig. 13). East and West orientations are the most problematic, since sun angles are too low to effectively block with overhangs or open blinds. Skylights in horizontal roofs can also be problematic since they receive excessive solar heat in summer and significantly less winter energy (when solar energy is most beneficial). Window sizes and locations should be uniquely considered for each building face, as well as the detailing of internal and external controls.



Facades and Enclosures, Building for Sustainability. Figure 12

Green walls at Musée Branly reduce heat gain in summer



Facades and Enclosures, Building for Sustainability. Figure 13

Critical facade elements for climate-responsive shading and glare control

Integral shading devices, the selection of reflective or low transmission glazing, or the introduction of “frit” patterns to reduce solar transmission, are the most common strategies for shading and glare control.

Unfortunately, low solar transmission glass will block useful gain in winter as well, and can also result in low daylight transmission. Many high reflectivity glass specifications inhibit daylighting, passive solar heating,

and clear views of the landscape. However, new developments in glazing support high visible transmittance (daylight and view) with low solar transmittance (solar heat) for buildings in hot climates. Just as low-e glazing is becoming the norm for improved thermal resistance in glazing assemblies, this high-visible, low-solar glazing must become the norm for all buildings where cooling loads dominate on the perimeter.

A number of new glazing materials are emerging that support seasonally dynamic management of solar heat and light. Electrochromic, prismatic, and holographic glazing assemblies are in development that would enable the seasonal or hourly separation of light and heat spectrums, and the depth and color of the beamed light, so that building spaces can be naturally conditioned with the least amount of solar overheating or glare. The combination of these heat and light redirection devices with exterior and interior shading devices enable the appropriate percentages of the solar light and heat to be independently tapped.

Exterior shading devices, overhangs, light shelves, blinds, fins, and shades are the most effective means of reducing solar load while maintaining visual connection with the landscape. Clearly, the shading devices must be designed for each orientation, and for each site's climate and latitude. At least six types of exterior devices can be explored: an overhang/light shelf, recessed windows that act as an "egg-crate" shading device, dynamic shades or awnings, exterior blinds, exterior green walls, and a second dynamic facade (Fig. 14).

Just as a baseball cap allows a clear view of the action without glare in the eye, overhangs or light shelves can allow for views without glare and solar overheating. While an overhang provides shade from high sun angles, it could act as a light shelf and reflect sunlight deeper into the space when positioned below the upper (transom) portion of the window wall. Fully retractable exterior blinds can support both seasonal shading and daylighting goals, and enable views to be fully enjoyed during those hours when the sun is away from the facade. Awnings offer seasonally and daily adjustable shading on all orientations, full views when the summer sun has passed, and a colorful gaiety to the building. Landscape materials are re-emerging as effective shading for roofs, walls, windows, and terraces, self-regulating their density of shade to match the seasons, and offering potential to manage storm water as well.

Green Walls = Energy Savings

In a 2000 controlled experiment at Tanz Greenhouse at the University of Toronto and a follow-up computer simulation study, Bass and Baskaran identify a 23% reduction in cooling load and a 20% reduction in the fan energy use, for an 8% reduction in total annual energy consumption, when a green wall was used to shade an exterior surface of the building, as compared to an unshaded surface (Tanz Greenhouse/Bass and Baskaran [28]).



Facades and Enclosures, Building for Sustainability. Figure 14

Exterior shading devices at the Chesapeake Bay Foundation Headquarters are designed to account for different building orientations

The line of last defense, *interior shading devices* are somewhat less effective than exterior shading at minimizing solar load (heat gain from the sun), since the energy is able to enter the building before it can be reflected. While roll-down mesh shades provide a level of solar heat and glare control, inverted horizontal venetian blinds provide heat and glare control as well as daylight contribution for sustainable buildings. Research at the Lawrence Berkeley National Laboratories demonstrates that exterior operable shading devices, and even interior shading devices, are more cost effective and energy/resource effective than highly tinted (low visible transmission) glass *in all regions of the USA*, from Los Angeles to Chicago to Miami [29]. The selection of interior blinds should maximize light reflection, typically with a white or metallic color; maximize light distribution in the room, typically through inverted blinds and carefully shaped blinds that act as effective light shelves; and provide adjustable slat angles to maximize view potential while also maximizing work-plane illuminance [30] (Fig. 15).

If they are combined with exterior shades in a multi-layered facade, interior shades and blinds can be dedicated to light diffusion and glare management, with fabric/color selection designed to minimize brightness contrast between the window and its surroundings. Coupled with operable exterior shading devices, interior blinds could also be selected to *absorb* solar heat for perimeter heat gain in winter months. As innovations in phase change materials emerge, it is

possible to imagine internal shutters or shades that absorb solar heat during the day, and then are reversed at night to radiate solar heat to the occupied space.

The most far reaching facade development in the past 10 years has been the development of *the double envelope*. When a second facade is introduced 1–3 m from the window wall, the intermediate space can be used to moderate the environment – heat, light, wind, noise, pollution, and other environmental stressors. These second facades house layers of shading, light redirection, air redirection, thermal load balancing, mechanical conditioning, and even dynamic levels of resistance to heat loss and gain. The research on the life cycle benefits of these innovations reveals significant variations in both costs and performance [31]. If double envelopes provide the greatest number of hours, days, and months without mechanical conditioning, they will be invaluable to carbon-neutral designers. The following section explores a number of enclosure solutions that use dimensioned “circulatory systems,” such as double envelope designs, to reduce environmental loads in buildings (Fig. 16).

Facades that Engineer Load Balancing and Mean Radiant Temperature Control

To work toward healthier buildings with higher levels of local/individual comfort, it is imperative that the thermal imbalances, *mean radiant imbalances*, and conductive losses in buildings be addressed (Fig. 17).



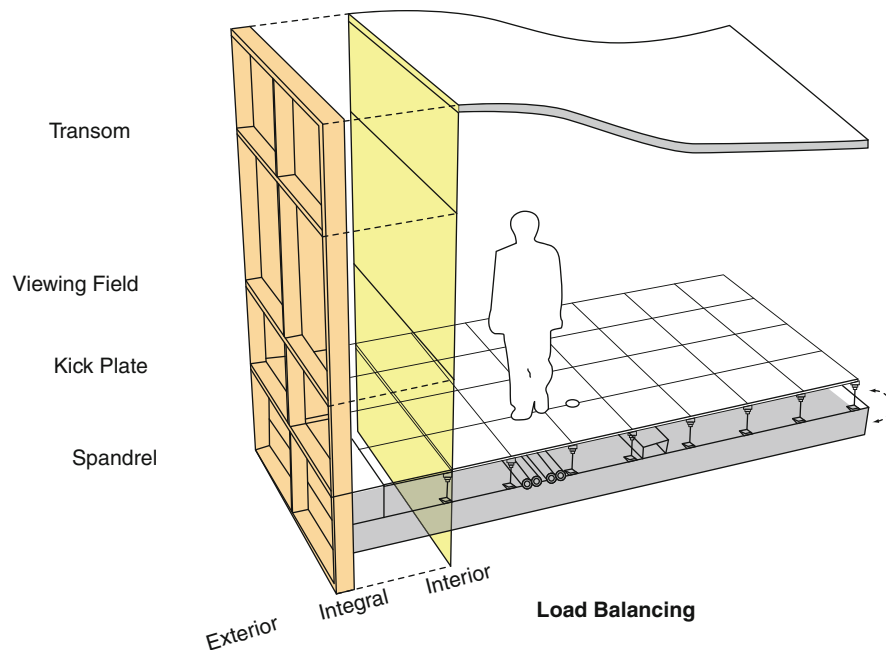
Facades and Enclosures, Building for Sustainability. Figure 15

The dynamic light redirection louvers in the Robert L. Preger Intelligent Workplace can be configured to increase daylight levels in the interior spaces while reducing glare at the perimeter. Interior venetian blinds have engineered w-sections that manage low and high angle sun differently, reducing direct glare while maximizing view and daylighting



Facades and Enclosures, Building for Sustainability. Figure 16

Light shelves in the double envelope combined with external overhangs provide wonderful daylight and views for the new Cambridge public library



Facades and Enclosures, Building for Sustainability. Figure 17

Critical facade elements for load balancing and mean radiant temperature control

The engineering team is equally responsible with the architect for ensuring that the building enclosure detailing fully resolves thermal bridges and radiant imbalances.

As the internal heat gains in buildings have increased with office automation, the disparity in temperature conditions between facades and interior spaces has grown. Surprisingly, some designers have removed or reduced perimeter conditioning on the assumption that the interior loads will compensate for perimeter losses. This convective exchange will not be effective, however, without specifying super-insulated facade assemblies and/or deliberate load balancing by mechanically moving “core” heated air or water through the window wall.

Both air flow windows and water flow mullion systems enable excess heat from the building core – heat from occupancy, lights, and equipment – to be effectively dissipated through the facade. By taking return air through the “glass duct” of a triple glazed air-flow window, core cooling loads are dramatically reduced and perimeter heating is almost eliminated through *core-to-perimeter load balancing*. Water mullions (thermally broken from the outdoors) can also use “waste” heat from the core to minimize loads. In

this system, waste heat from building cooling or power generating systems can be circulated through the facade to eliminate perimeter heating requirements and radiant imbalance, while allowing an increase in building periphery for views and light at every workstation. Indeed, the building facade could be seen as the natural dissipater of energy, a “circulatory system” resembling that of a healthy human – with appropriate surface to volume ratios. Given the constant increases in internal loads today, there is a real justification for increasing the periphery of buildings with the facade designed as an integral part of the mechanical system.

In addition to core and perimeter load differences, there are also significant load imbalances at different facade orientations, typically driven by solar load differentials (Fig. 18). One facade mechanical system of note is the use of double envelopes to provide *north-south-east-west load balancing* in climates where solar loads are significant and beneficial. The Occidental building in Niagara Falls is one of the earliest examples of a double envelope construction. When solar energy is received on the east, a natural convective loop of solar-heated air wraps the entire building. This continues throughout the day to eliminate simultaneous heating and cooling, to maximize passive solar



Facades and Enclosures, Building for Sustainability. Figure 18

In addition to radiant ceiling panels, radiant water mullions are integrated into the facade to balance temperatures year round in the Intelligent Workplace using water for energy-efficient load balancing

contributions to the heating load, and to ensure excellent mean radiant conditions. In summer, the double envelope is vented to the outside, and precautions are taken for fire protection. The Occidental building uses 70% less heating and cooling energy than a conventional office building in upstate New York [32].

A third strategy for reducing or eliminating the need for mechanical cooling is the use of *day-night load balancing* strategies. Night ventilation of a building's structural mass can successfully reduce or eliminate cooling loads in cooler climates, or climates with day-night temperature swings. Combined with the careful monitoring of dew point temperatures to avoid overcooling and condensation, nighttime cooling of a building's mass can provide adequate capacitance to absorb an entire day of internal gains. Thermal mass can be achieved through "heavy" or high-capacitance building walls and floors, or through the strategic introduction of phase change materials (PCM) to absorb and radiate heat as designed. A corollary to nighttime ventilation is the use of off-peak nighttime electricity to generate chilled water and ice or to chill PCMs for daytime cooling. This "ice storage" cooling strategy can be effectively bundled with nighttime ventilation of a building's thermal mass to absorb daytime overheating loads without peak electricity demands. Phase change materials also offer the potential for storing daytime solar heat for nighttime use, a strategy integrated into a number of PassivHaus projects [33].

For opaque enclosure surfaces, the value of high reflectivity surfaces in all overheated climates has led to a new era of "cool roof" design [34]. The combination of high reflectivity and night venting of enclosure mass in climates with day-night temperature swings has the greatest potential for natural comfort and low energy cooling [35].

Cool Roof = Energy Savings + Peak Demand Reduction + Extended Roof Life

In a 2002 multiple building controlled experiment in Florida, Parker et al. of the Florida Solar Energy Center identify 17.8–24.9% annual cooling energy savings and a 28.5–35.5% peak cooling load reduction from highly reflective white metal, white barrel tile, and white flat tile roofing; 9.5% annual cooling energy savings and a 12.9% peak cooling load reduction from terra cotta

tile roofing; and 3.2% annual cooling energy savings and a 17.2% peak cooling load reduction from white asphalt shingles, as compared to dark asphalt shingles, on a single-family house (Florida Sustainable Energy Center/Parker et al. [35]).

The final load-balancing strategy for reducing heating and cooling loads in buildings is to take advantage of ground and aquifer temperatures through *ground-coupling*. Ten feet below ground, and in underground aquifers, the temperatures are surprisingly stable at year-round (inter-seasonal) averages of approximately 12–18°C (55–65°F). Given excellent temperatures for radiant, water and air-based cooling, high-performance building designers have tapped into this natural cooling potential through earth sheltered construction, large diameter "cool tubes," and aquifer or ground sourced chilled water loops that serve radiant, fan coil, or heat pump systems. Ground coupling significantly impacts foundation and flooring design, and might impact the facade "circulatory" systems as well (Fig. 19).

The benefits of active (mechanically integrated) load balancing to resolve mean radiant imbalances – from core to perimeter, from N/S to E/W, from day to night, and inter-seasonally through ground coupling – are many and are as follows: improved individual comfort, productivity, and health [36, 37]; reduced building material degradation from thermal differentials and condensation; significant energy conservation; and the potential to replace mechanical heating or cooling first costs and operating costs with high-performance building enclosures [38] (Fig. 20).

Radiant Cooling = Individual Productivity + Energy Savings

In a 1999 controlled field experiment and simulation study, Takehito et al. identify a measured 22% increase in speed and a 1.5% improvement in accuracy on simple tasks among women subjects and a simulated 10% HVAC energy savings in the Tokyo climate from a radiant ceiling cooling system as compared to conventional ceiling-based air conditioning (Imanari et al. [36]).



Facades and Enclosures, Building for Sustainability. Figure 19

“Cool tubes” beneath the greenhouses at Phipps Conservatory provide free cooling

Night Ventilation Cooling = Individual Productivity

In a 2003 meta-analysis study, Seppänen et al. identify a productivity increase equivalent to 0.39 h of work per day (4.9% for an 8-h workday) due to nighttime ventilation cooling of thermal mass, a very energy-efficient method of reducing daytime indoor temperatures by using nighttime air to cool a building’s structure and furnishings (Seppanen et al. [37]).

Geothermal Heat Pump = Energy Savings + First Cost Savings

In a 1998 field study of Maxey Elementary School in Lincoln, Nebraska, Shonder et al. identify an average 47% reduction in annual HVAC energy use with a geothermal heat pump system as compared to the alternatives of VAV systems with air- or water-cooled chillers, as well as a first cost savings of \$1.80/sf. Geothermal heat pump systems were found to have the lowest first cost, operational cost, and life cycle emissions of the three systems (Maxey Elementary School/Schonder et al. [38]).

Facades that Integrate Passive and Active Solar Heating, Cooling, and Power

The most direct use of solar energy in buildings is for *passive solar heating* (Fig. 21). The design of building enclosures to support direct gain through high solar transmission glass can effectively offset over 50% of the heat loss of buildings in many climates. Indeed, it is critical that glazing be carefully selected to ensure high solar transmission in cold climates to enable passive solar heating. In locations where direct gain would cause excessive daytime overheating or glare, the design of indirect gain passive solar facades – glazed mass walls or Trombe walls [39], water walls, or phase change walls – can strategically collect solar heat for nighttime heating. Recent developments in phase change materials that can be embedded in common construction or finish materials will help ensure that adequate solar heat storage exists to shift daytime solar gain to nighttime loads. Each of these solar heating strategies can be fully passive, utilizing no mechanical energy to collect, store, or distribute solar heat while providing potential health benefits to building occupants [40] (Fig. 22).

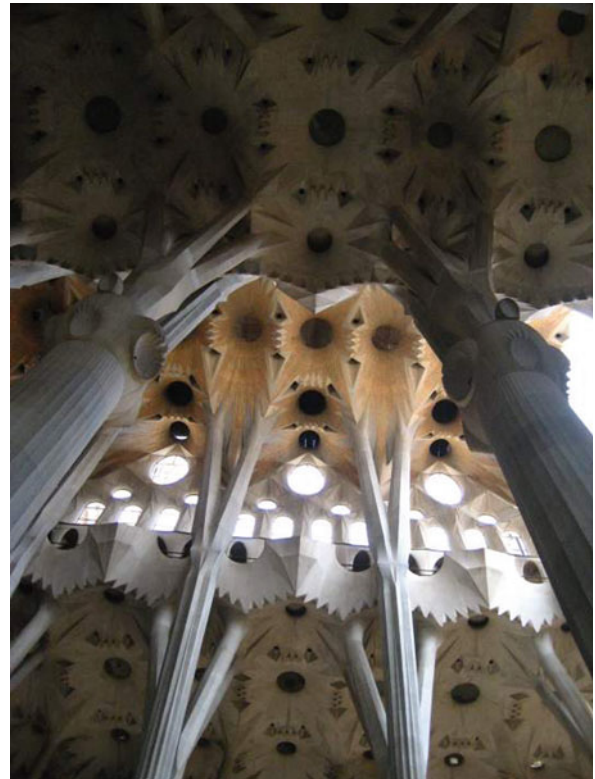
Sunlight = Health

In a 2005 building case study of Inha University Hospital in Korea, Choi identified a 41% reduction (3.2 days) in average length of stay among gynecology patients in brightly daylighted rooms (317 lx average), as compared to those in dull rooms, in the spring, and an average 26% reduction (1.9 days) in average length of stay among surgery ward patients in bright rooms, as compared to those in dull rooms, in the fall. Across all seasons, the average daylight illuminance in bright rooms was 317 lx, compared to 190 lx in dull rooms (Inha University Hospital/ Choi [40]).

The solar heat gained in building atria or double envelope facades can also be actively transferred to a range of storage media, from rock storage to phase change storage to underfloor air systems with thermal mass. *Active solar hot water* collectors integrated into building roofs or facades can also transfer solar heat to architectural or mechanical heat exchangers. The effective integration of solar heat gain with the ventilation air systems in buildings can significantly reduce heating loads while supporting greater ventilation rates for human health.

While solar thermal collectors can provide excellent hot water temperatures for heating, *concentrating solar thermal* collectors can generate the extreme high temperatures needed for cooling and power generation. Bundled with the latest in absorption chillers and steam power generators, a new generation of solar-evacuated tube and concentrating collectors are beginning to transform building enclosures to help meet the growing electrical demands in buildings for air conditioning (Fig. 23).

Finally, integrating *photovoltaic solar* (PV) elements into the building curtain wall or roof can generate on-site power for feeding into the grid or for immediate use in the building. For maximum efficiency, PV energy could be used directly as DC electricity (without transformer losses to AC) to meet the power demands of fans and air conditioners, plug loads, and dynamic facade elements – from operable windows to light shelves and shading devices. Recent thin film developments have significantly improved the performance of building integrated photovoltaics (BIPV), supporting

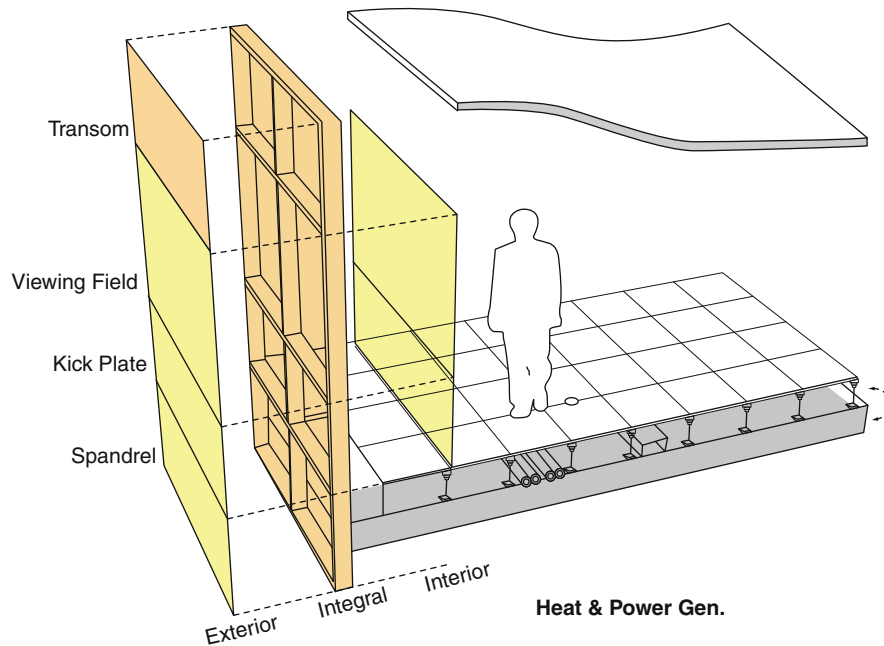


Facades and Enclosures, Building for Sustainability.

Figure 20

Architecturally articulated thermal mass at Gaudi's Sagrada Família minimizes daily temperature swings by night ventilation through architecturally expressive openings

a range of design choices regarding shape, orientation, transparency, and color. For windows, new PV glazing assemblies can provide effective shading, daylight distribution, and a competitive power source. The next generation of building windows may absorb solar energy in almost transparent coatings and carry the electricity to window mullions for transport to the grid. For roofs and walls, the early PV assemblies that were rigidly sandwiched behind glass have been replaced by newer flexible PV materials that can replace common building materials, with 3–10-year life-cycle justifications. The vast areas of roof and facade that absorb or reflect long hours of sunshine can be turned into an asset even in hot climates, with building integrated photovoltaics playing a major role in generating distributed power, eliminating the source and



Facades and Enclosures, Building for Sustainability. Figure 21

Critical facade elements for passive and active solar heating, cooling, and power



Facades and Enclosures, Building for Sustainability. Figure 22

Interior and exterior views of the south facade at Fraunhofer Institute in Freiburg illustrate the integration of active and passive solar strategies

distribution inefficiencies of conventional power production for greatly reduced carbon impacts.

Wind power generation can also be integrated with building facades and roofs. A new generation of micro-

turbines may create the parapets of future buildings, and large wind farms may occupy the upper floors of skyscrapers, as proposed in early solutions for the World Trade Center by Battle Engineering [41].



Facades and Enclosures, Building for Sustainability. Figure 23

Concentrating solar thermal collectors on the roof of the Intelligent Workplace produce steam in the summer to power absorption chillers to cool the space

Photovoltaic System = Energy Savings

In a building case study of Bren Hall, a LEED Platinum-rated building at the University of California, Santa Barbara (UCSB), Aigner identified that the building's 47 kW rooftop photovoltaic system reduces purchased electricity consumption by 7–10% annually, at a first cost of \$240,000, for an ROI of 5% (Bren Hall/ Aigner [42]).

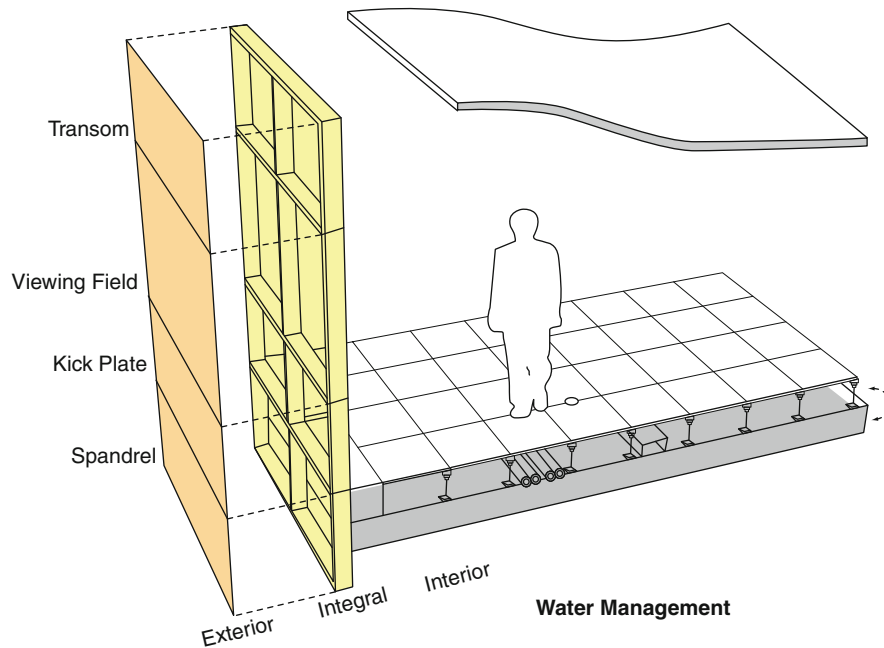
Facades Designed to Manage Water

A fundamental role for high-performance building enclosures is the management of water. It is imperative that roofs and facades effectively *manage rainfall* through material selection and articulated detailing from top to bottom (Fig. 24). While gargoyles and expressive cornices that were once the solution to ensure heavy rains were shed away from the building, modern facades rely far too heavily on the caulked joint – often with poor results as the facades age (Fig. 25).

A lesser understood role for building facades is to *manage the vapor migration* that might lead to

condensation on embedded surfaces that are below dew point temperature. This is typically achieved through the design of continuous vapor barriers on the “warm side of the facade” to stop the humidity from migrating into the wall. Detailing in design and construction supervision is critical, especially in complex climates where humidity and cold temperatures may both be from the inside or outside. In addition, moisture that is trapped in construction materials must be released naturally as well, typically through a vapor porous skin or vented air spaces [43]. *Rain screens* are an innovative solution to managing both rain and vapor migration. Rain screens introduce an uncaulked and vented second skin for the facade – typically of ceramic tile or metal or wood panels – that sheds the rain and utilizes the vented air space to eliminate moisture buildup.

As we embrace a greater set of sustainability challenges, building enclosures also need to collect rainwater today. As we face declining sources of fresh water, and increasing storm runoff challenges, the importance of designing building enclosures to *catch and store water* is of growing importance (Fig. 26).



Facades and Enclosures, Building for Sustainability. Figure 24
Critical facade elements to manage rainwater and moisture migration



Facades and Enclosures, Building for Sustainability. Figure 25
On the *left*, water sculpture by Herbert Dreisetl playfully manages storm water, enticing visitors to return in the rain to watch. On the *right*, rainwater capture provides acoustic and visual richness in the Dockside Green development by Joe Van Bellingham

Green Roof = Energy Savings + Peak Load Reduction + Storm Water Retention

In a 2002 building case study of the Chicago City Hall, the city of Chicago identified annual cooling energy savings of 0.02 kWh/square foot and annual heating energy savings of 0.02 therms/square foot, as well as a 70% reduction in storm water runoff, from installation of a 20,000 sf green roof (Chicago City Hall/ City of Chicago [44]).

Green Roof = Energy Savings + Storm Water Retention

In a 2003 building case study of the 901 Cherry Offices of Gap, Inc. in San Bruno, CA, Green Roofs for Healthy Cities identify a 100% reduction in storm water runoff (7.54 gal/sf of roof area) and a simple payback of 11 years from energy cost savings alone (equivalent to 7% annual energy savings in a baseline building) due to installation of an extensive green roof (901 Cherry (GAP)/ Green Roofs for Healthy Cities [45]).

Facades that Maximize Enclosure Life

High-performance building enclosures will need to *use materials sustainably* (Fig. 27). Wherever possible, existing enclosures should be preserved to protect their “embodied” energy, or built of 100% recycled content materials. If virgin materials are used, their “chain of custody” should be known in relation to energy, carbon, rare materials, and toxicity. Toxic materials should be avoided in all assemblies, or designed for retrieval during disassembly.

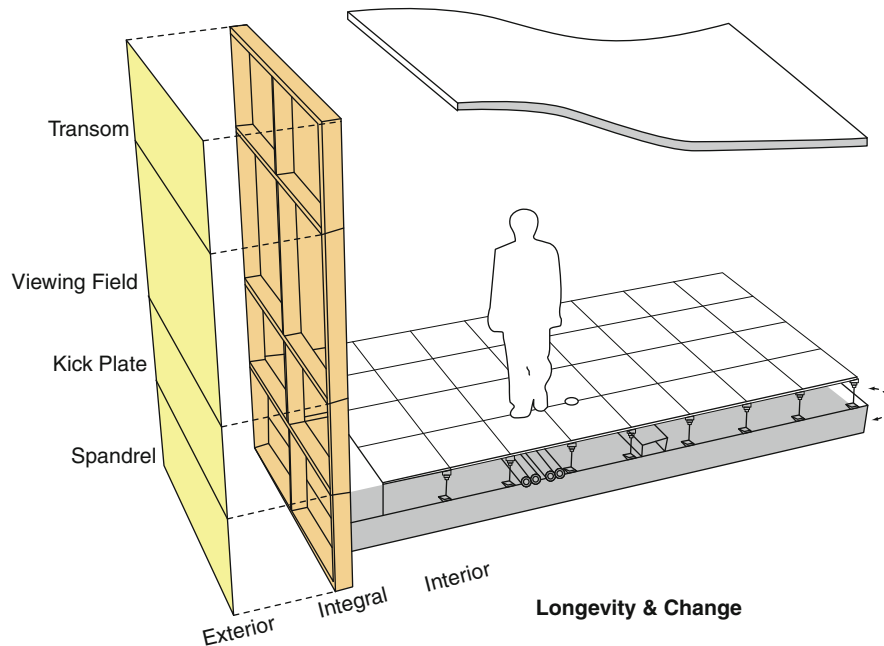
Enclosure assemblies should be designed for their longevity, for natural *weathering and effective maintainability*. Most critically, enclosures should be designed and built for *cherish-ability*, the desire on the part of owners and the community to want to preserve the building rather than tear it down. This is typically ensured by the use of quality materials, assembled for longevity, and a high level of craftsmanship (Fig. 28).

Enclosure design must also address the needs for *human safety* in the face of manmade and natural disasters. This includes structural integrity in the face of earthquakes, hurricanes, and cyclones as well material integrity in the face of fires, floods, and blasts.



Facades and Enclosures, Building for Sustainability. Figure 26

Green roofs at the California Academy of Sciences capture storm water while providing a learning environment for landscape and habitats



Facades and Enclosures, Building for Sustainability. Figure 27

Critical facade elements to maximize enclosure life

Design for disaster protection should be about human safety not building protection, ensuring adequate egress and adequate safe harbors. Design for human safety must also ensure adequate access to water, breathing air, and light, even in the face of power outages – a redundancy that can be best achieved by sustainable facade design that fully incorporates natural conditioning.

At the same time, building enclosures need to be *designed for change* – for expansion or contraction, for changes in use or function. Change can be supported through building generosity, floor-to-floor heights, size of openings, and level of building articulation. Change can be supported by modularity, with a kit of parts that anticipates the evolution of scale and function over the life of the building. Change can also be accommodated by *design for disassembly*, by which the building and/or its enclosure can be relocated to a second or third life. All high-performance enclosure components, such as window and door assemblies, water management systems, and ornamental parapets, should be designed to be recovered as the “value-added” assemblies that they are well beyond the value of the raw materials (Fig. 29).

Construction Waste Management = Salvage/Waste Savings

In a 2004 building case study of the renovation of the Boston Scientific Company, Inc. office building, the Institutional Recycling Network identified a 32% reduction in disposal costs due to a comprehensive waste management process, compared to the cost of renovation with conventional landfill disposal of material waste. A site separation process allowed materials to be disposed of at lower costs than co-mingled materials and resulted in a 92% recycling rate by weight (Boston Scientific Company/ Institution Recycling Network [46]).

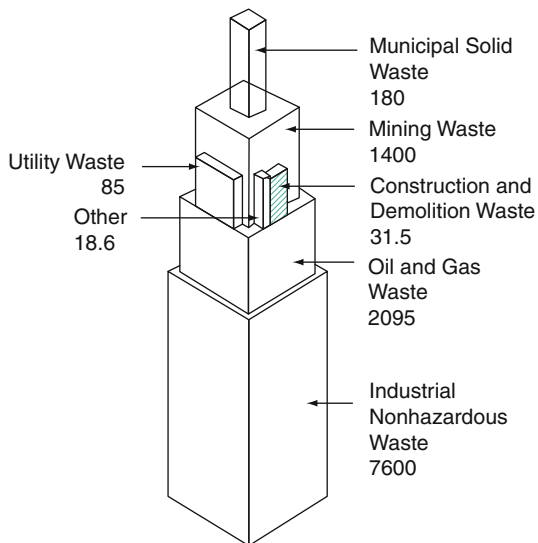
Construction Waste Management = Salvage/Waste Savings

In a 2000–2001 building case study of the demolition of the 133,000 square foot Hume Residence Hall on the campus of the University of Florida, Guy and Strong identify new revenues, reduced disposal costs, and cost avoidance from substitutions of reclaimed

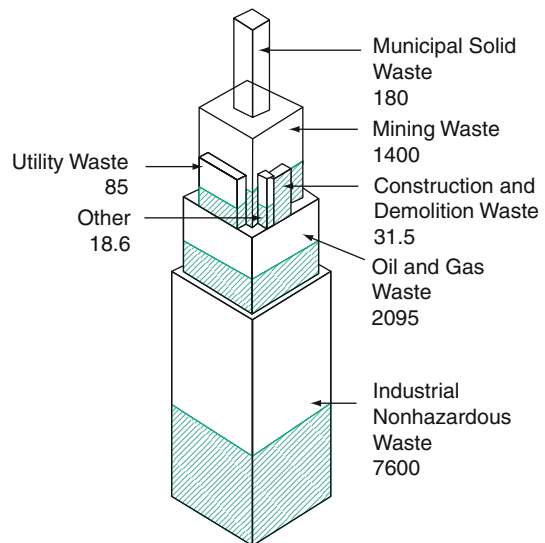


Facades and Enclosures, Building for Sustainability. Figure 28
Historical structures were designed and built to last a lifetime

Annual Waste Breakdown for the U.S. by Source



EPA / EPRI ESTIMATE
(in Million Tons)



Building Industry related wastes
(40% of each segment)

CENTER FOR BUILDING PERFORMANCE &
DIAGNOSTICS / ABSIC, VOLKER HARTKOPF
ESTIMATE

Facades and Enclosures, Building for Sustainability. Figure 29
Forty percent of the total waste stream can be attributed to the construction industry

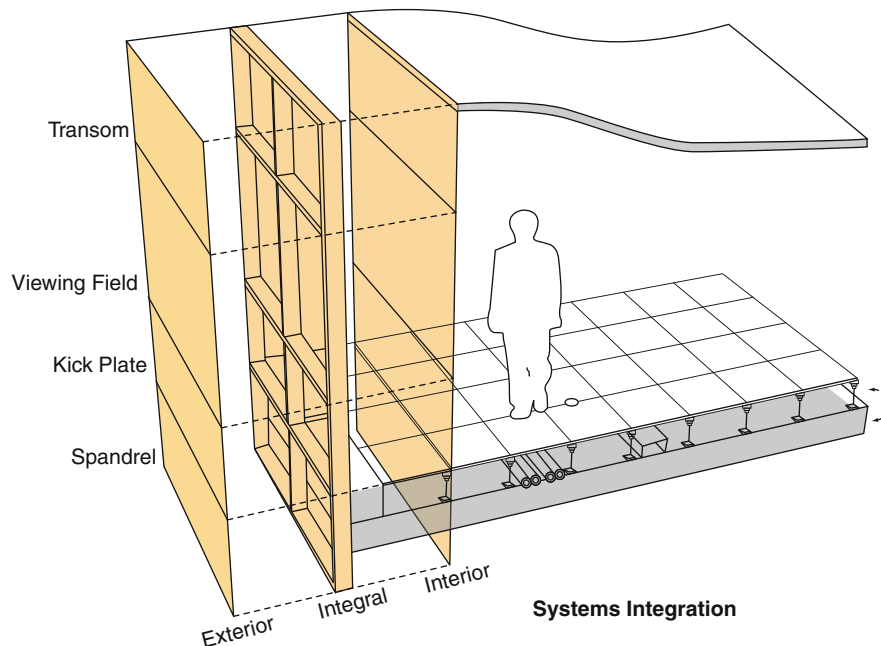
materials, achieved through a process of salvage and source separation of “debris” building materials. A comprehensive C&D debris management process resulted in a *net savings* of 4% compared to the cost of conventional demolition with extensive landfill disposal (University of Florida Hume Residence Hall/Guy and Strong [47]).

Future Directions: Facade Innovation Through Systems Integration

In times of limited resources and limited building budgets, some of the most innovative high-performance facades are achieved through systems integration (Fig. 30). At a very minimum, high-performance sustainable building facades will have to be collaboratively designed to:

- Integrate structure and enclosure to ensure heat loss/heat gain control, provide shading and glare control, and support daylighting.
- Integrate fire and enclosure to support load balancing innovations.
- Integrate HVAC and enclosure to support natural ventilation, load balancing, and passive and active solar energy utilization.
- Integrate interior systems and enclosure systems for access to nature, daylighting, natural ventilation, shade and glare control, and passive solar heating.
- Ensure design for enclosure life and design for change without waste.

When expert disciplines work collaboratively early in the design process, a range of innovations are generated – structural systems that offer shading and light redirection; enclosures that are integral with the mechanical duct or piping; fire egress that provides shade and access for maintenance; ornament that manages water; and expressions of function, culture, or ambition that serve a critical purpose for ensuring the highest level of indoor environmental quality and environmental sustainability. Sustainable building enclosures are the deliverables of integrated design teams and integrated building delivery processes that in turn are pivotal to achieving the highest level of health, productivity, energy, and environmental benefits (Fig. 31).



Systems Integration

Facades and Enclosures, Building for Sustainability. Figure 30
Critical facade elements for innovation through systems integration

Whole Building = Energy Savings

In a 2000 building case study of BRE's Environmental office building in Garston, UK, Fisher et al. identify a 66% annual energy savings compared to a typical type III UK office building due to the use of thermal mass, natural ventilation, passive and low energy cooling, daylighting, high-performance electric lighting, and on-site power generation (BRE Environmental Building/Fisher et al. [48]).

Floor-Based Ventilation + Increased Outside Air = Health

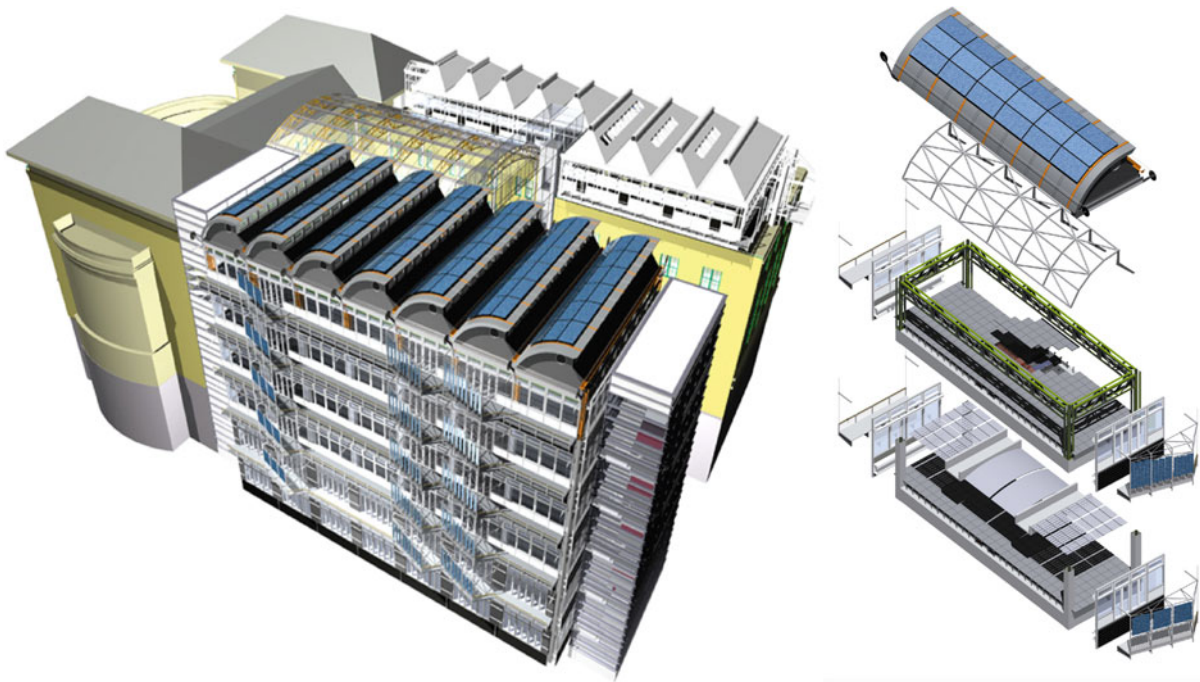
In a 2000 multiple building study of 39 schools in Sweden, Smedje and Norback identify a 69% reduction in the 2-year incidence of asthma among students in schools that received a new displacement ventilation system with increased fresh air supply rates and dedicated exhaust, as compared to students in schools that did not receive a new ventilation system (Elementary School/Smedje and Norback [49]).

High-Performance Building = Energy Savings + Rent Premium

In a 2004 building case study of the Phillips Eco-Enterprise Center in Minneapolis, Brinkema identifies \$0.31/sf energy savings and a 5–10% lease premium for the building owner (a financial gain of \$0.62/sf) due to high-performance building design that incorporates passive solar and geothermal heating and cooling, daylighting and high efficiency electric lighting, salvaged building materials, nontoxic paints and finishes and landscaping with native plants (Philips Eco-Enterprise Center/ Brinkema [50]).

High-Performance Building = Energy Savings + Individual Productivity

In a 1992 building case study of ING Bank in Amsterdam, Bill Browning of Rocky Mountain Institute identifies a 92% reduction in primary energy consumption and a 15% reduction in employee absenteeism compared



Facades and Enclosures, Building for Sustainability. Figure 31

The proposed "Building As Power Plant" at Carnegie Mellon University attempts to integrate all building systems, assemblies, and components to export unneeded renewable energy while providing optimum comfort for the occupants

to the bank's former headquarters, due to high-performance design strategies including daylight, a narrow floor plan that allows landscaped views for every occupant, passive solar conditioning, co-generation, and the use of heat exchangers (ING Bank/ Browning [51]).

Bibliography

1. Mahdavi A, Brahme R, Mathew P (1996) The "LEK"-concept and its applicability for the energy analysis of commercial buildings. *Build Environ* 31(5):409–415
2. Wilson E (1984) *Biophilia*. Harvard University Press, Cambridge
3. Kellert S, Wilson E (1993) *The biophilia hypothesis*. Island Press, Washington, DC
4. Kellert SR (2005) *Building for life: designing and understanding the human-nature connection*. Island Press, Washington, DC
5. Kellert S, Heerwagen J, Mador M (2008) *Biophilic design: the theory, science and practice of bringing buildings to life*. Wiley, Hoboken
6. NKB (1991) *Indoor climate – air quality*, Nordic Committee on Building Regulations. NKB Publication No 61E, June 1991
7. CBPD (n. d.) *eBIDS – Energy building investment building support*. <http://cbpd.arc.cmu.edu/ebids/pages/home.aspx>. Accessed 7 Feb 2011
8. Ulrich RS (1984) View through a window may influence recovery from surgery. *Science* 224(4647):420–421
9. Mendell MJ (1991) Risk factors for work-related symptoms in northern California Office workers. Indoor Environment Program Lawrence Berkeley Lab, CA, Unpublished doctoral dissertation, University of California, CA Department of Health Services, and NIOSH, Sponsoring Org. USDOE, Washington, DC, 1991
10. Heschong Mahone Group (2003) *Windows and offices: a study of office worker performance and the indoor environment*. CEC PIER Report, Oct 2003
11. Loftness V (1994) Major recommendations for action: forrestal and germantown building evaluation team. Field Study. Center for building performance and diagnostics, Carnegie Mellon University, Pittsburgh, 17–18 Oct 1994
12. Moore RC, Marcus CC (2008) Healthy planet, healthy children: designing nature into the daily spaces of childhood. In: Kellert S, Heerwagen J, Mador M (eds) *Biophilic design: theory, science and practice of bringing buildings to life*. Wiley, Hoboken, pp 153–203
13. US DOE (2009) *2008 Building energy data book*. Prepared for the Building Technologies Program Energy Efficiency and Renewable Energy, U.S. Department of Energy. By D&R International. Under contract to National Energy Technology Laboratory, 2006 Energy Consumption
14. Loftness V, Aziz A, Choi J (2008) Energy savings and performance gains in GSA buildings: seven cost-effective strategies. GSA Public Buildings Service, Washington, DC
15. Rubenstein F, Jennings J, Avery D, Blanc S (1998) Preliminary results from an advanced lighting controls testbed. In: 1998 illuminating engineering society of north america annual conference, San Antonio, TX, USA, 10–12 Aug 1998
16. EPRI (1997) Performance evaluation of energy efficient lighting and office technologies in New York City – Final Report. Electric Power Research Institute, EPRI, Palo Alto, June 1997
17. Boyce PR, Beckstead JW, Eklund NH, Strobel RW, Rea MS (1997) Lighting the graveyard shift: the influence of a daylight-simulating skylight on the task performance and mood of night-shift workers. *Light Res Technol* 29(3):105–134
18. Preziosi P, Czernichow S, Gehanno P, Hercberg S (2004) Workplace air-conditioning and health services attendance among French middle-aged women: a prospective cohort study. *Int J Epidemiol* 33(5):1120–1123
19. Emmerich SJ, Persily AK, Stuart Dols W, Axley JW (2003) Impact of natural ventilation strategies and design issues for California applications, including input to ASHRAE Standard 62 and California Title 24. NISTIR 7062, 2003
20. Rowe D (2002) Pilot study report: Wilkinson building. The University of Sydney, Sydney
21. Hannula M, Niemela R, Rautio S, Reijula K (2000) The effects of indoor climate on productivity. In: *Proceedings of healthy buildings*. Helsinki, pp 659–664
22. Loftness V, Baird N, Snyder M (2011) The triple bottom line of green roofs: assessing the life cycle economic, environmental and human benefits of green roof. Carnegie Mellon University, Pittsburgh
23. Kosareo L, Ries R (2007) Comparative environmental life cycle assessment of green roofs. *Build Environ* 42(7):2606–2613
24. Gangnes D (2007) Seattle green roof evaluation project final report. Magnusson Klemencic, Seattle: Seattle.gov
25. Currie BA (2005) Air pollution mitigation with green roofs using the UFORE model. Ryerson University, Toronto
26. Onmura S, Matsumoto M, Hokoi S (2001) Study on evaporative cooling effect of roof lawn gardens. *Energy Build* 33(7):653–666
27. Stec WJ, van Paassen AHC, Maziarz A (2005) Modelling the double skin façade with plants. *Energy Build* 37(5):419–427
28. Bass B, Baskaran B (2003) Evaluating rooftop and vertical gardens as an adaptation strategy for urban areas. National Research Council Canada, Report No. NRCC-46737, Toronto, Canada, pp 74–86
29. Winkelmann FC, Lokmanhekim M (1981) Life-cycle cost and energy-use analysis of sun-control and daylighting options in a high-rise office building. Report No. LBL-12298, University of California, Lawrence Berkeley Laboratory, Berkeley
30. Koster E (1998) *Natuur onder architectuur = architecture for nature*: IBN-DLO Wageningen, architect stefan Behnisch. Schuyt, Haarlem
31. Oesterle E, Lieb RD, Lutz M, Heusler W (2001) Double skin facades: integrated planning: building physics, construction, aerophysics, air-conditioning, economic viability. Munich, Prestel, London/New York
32. Bazjanac V (1980) Architectural energy analysis. *Prog Architect* 61(4):98–101

33. Feist W (1999) Das Passivhaus. CF Mueller, Heidelberg
34. Akbari H, Levinson R (2008) Evolution of cool-roof standards in the US. In: Santamouris M (ed) *Advances in building energy research (ABER)*, vol 2. Earthscan, London, pp 1–32
35. Parker D, Sonne J, Sherwin J (2002) Comparative evaluation of the impact of roofing systems on residential cooling energy demand in Florida. In: *Proceedings of ACEEE 2002 summer study*, American Council for an energy efficient economy, Washington DC, August 2002
36. Imanari T, Omori T, Bogaki K (1999) Thermal comfort and energy consumption of the radiant ceiling panel system. Comparison with the conventional all-air system. *Energy Build* 30(2):167–175
37. Seppanen O, Fisk WJ, Faulkner D (2003) Cost benefit analysis of the night-time ventilative cooling in office building. Lawrence Berkeley National Laboratory, Berkeley. LBNL Paper LBNL-53191, 1 June 2003
38. Shonder JA, Martin MA, Hughes PJ, Thornton J (2000) Geothermal heat pumps in K-12 schools: a case study of the Lincoln Nebraska, schools. ORNL/TM-2000/80, Oak Ridge National Laboratory, Oak Ridge
39. Hastings R, Wall M (2007) Sustainable solar housing: exemplary buildings and technologies. International Energy Agency. Solar Heating and Cooling Programme, IEA energy conservation in buildings and community systems Programme, vol 2. Earthscan, London\Sterling
40. Choi J (2005) Study of the relationship between indoor daylight environments and patient average length of stay (ALOS) in healthcare facilities. Unpublished Master's Thesis, Department of Architecture, Texas A&M University. College Station, TX, USA
41. Ragheb M (2008) Wind turbines in the urban environment. Wind power systems: Harvesting the wind. University of Illinois at Urbana-Champaign, 30 Dec 2008. <https://netfiles.uiuc.edu/mragheb/www/NPRE%20475%20Wind%20Power%20Systems/Wind%20Turbines%20in%20the%20Urban%20Environment.pdf>. Accessed 7 Feb 2011
42. Aigner DJ (2002) Bren Hall – a living laboratory. *Solar Today*, 28–31 July/Aug 2002
43. Lstiburek J, Carmody J (1996) *Moisture control handbook: principles and practices for residential and small commercial buildings*. Wiley, New York
44. City of Chicago (2000) Department of environment, extensive green roofs fact sheet
45. Green Roofs for Healthy Cities (2003) Green roof awards of excellence – profile of award winners. In: *The First annual green roof infrastructure conference*. Illinois, Chicago, 29–30 May 2003
46. Institution Recycling Network (2004) Boston Scientific Company, Inc. (BSCI) Office Park Renovation, Marlborough, MA, 2004. <http://www.wastemiser.com/CS-BSCI-10-06.pdf>. Accessed 18 Nov 2007
47. Guy GB, Strong KS (2001) Deconstructing Hume residence hall. Report of the Powell Center for Construction and Environment, University of Florida, Gainesville
48. Fisher J, NiRiain C, MacKenzie F, Littler J (2000) BRE's environmental building: energy performance in use. In: *CIBSE Conference Paper*, Dublin
49. Smedje G, Norbäck D (2000) New ventilation systems at select schools in Sweden – effects on asthma and exposure. *Arch Environ Health* 55(1):18–25
50. Brinkema C (2004) EIN and real estate development. EIDC Roundtable. Fall, Vancouver
51. Browning W (1992) NMB bank headquarters: the impressive performance of a green building. *Urban Land*, June 1992, pp 23–25

Fallout Radionuclides and the Study of Erosion and Sedimentation

DESMOND E. WALLING

Department of Geography, College of Life and Environmental Sciences, University of Exeter, Exeter, UK

Article Outline

Glossary

Definition of the Subject

Introduction

A Brief History of the Use of Fallout Radionuclides in Erosion and Sedimentation Studies

Fallout Sources and Inputs

Fallout Radionuclides in the Soil and Soil Inventories
Using Fallout Radionuclides as Tracers to Investigate Soil and Sediment Redistribution within the Landscape

Field and Laboratory Techniques

Some Examples of the Application of Fallout Radionuclides in Erosion and Sedimentation Investigations

Future Directions

Bibliography

Glossary

Activity The concentration of a given radionuclide in a soil or sediment per unit mass, also referred to as the mass activity density.

Erosion The mobilization of soil and sediment from its original location or a temporary store within the landscape by a range of agents, including water and wind.

Fallout radionuclides Radionuclides that enter the terrestrial and aquatic environments as fallout.

Fluvial system Landscape systems associated with streams and rivers and driven by rainfall and runoff.

Gamma spectrometry Measurement of the gamma-emitted radioactivity in soils and sediment, using a gamma spectrometer.

Gross erosion The total amount of soil mobilized by erosion within a given area.

Half-life The time taken for a radionuclide to decay to the point where its activity is half the original activity.

Inventory An alternative term for the areal activity density, which represents the total radionuclide content of a soil or sediment deposit per unit surface area.

Net erosion The total amount of eroded sediment transported out of a given area. This represents the gross erosion within that area minus the deposition occurring within that area.

Off-site impacts The impacts of soil erosion associated with the transport of eroded soil away from the original point of erosion, for example, by streams and rivers.

On-site impacts The impacts of soil erosion associated with the point where erosion occurs and including reduction in soil depth, loss of productivity, and destruction of agricultural land.

Sediment budget A conceptual model of the sources, transfers, sinks, and outputs of sediment for a catchment or river basin. The model may quantify mobilization, transfer, storage, and output.

Sediment delivery ratio This represents the ratio of the amount of sediment leaving a field or catchment to the total amount of sediment mobilized by erosion within that area. It provides a measure of the efficiency of sediment transfer out of the system considered. The sediment delivery ratio will decrease as deposition increases.

Sediment Material transported or deposited within fluvial, aeolian, and other erosional systems.

Sediment source fingerprinting Use of a range of biogeochemical properties to characterize or fingerprint different potential sediment sources. The properties of transported or deposited sediment are compared with those fingerprints, in order to determine its source.

Sediment source The point, place, or area from which sediment was mobilized.

Sedimentation A term frequently used to refer to the study of sediment transport in fluvial systems, including sediment mobilization, transport, and deposition, but also used in a more narrow sense to refer to deposition of sediment in sinks including river channels and floodplains, lakes and reservoirs, and estuaries and coastal environments.

Definition of the Subject

Soil erosion is an important problem in many areas of the world, where it reduces soil productivity, destroys agricultural land, and can threaten food security and the sustainable management of soil resources. The on-site effects of soil erosion are coupled with off-site effects, which relate to the downstream or downwind impact of the eroded soil. The resulting transport of increased amounts of fine sediment, particularly in streams and rivers, and the deposition of this sediment in downstream sinks gives rise to a range of important environmental problems. These include the degradation of water quality and aquatic habitats. Concern for the sustainable management of both soil resources and catchments and river basins has directed attention to the need to develop effective erosion and sediment control strategies. Designing and implementing such strategies requires an improved understanding and quantification of the mobilization, transfer, and deposition of sediment within the landscape, so that mitigation measures can be targeted to optimize their effectiveness. Existing measurement techniques for assembling data on erosion and sedimentation within the landscape face many limitations, and the potential for using fallout radionuclides to assemble the information required has attracted increasing attention over the past 30 years. Fallout radionuclides offer unique opportunities to document sediment mobilization and deposition rates and to trace the movement of fine sediment through the landscape. This entry reviews the important needs for information on erosion and sedimentation and the history of the application of fallout radionuclides in this field, describes and explains the basis for using fallout radionuclides in erosion and sedimentation investigations, and provides several different examples of their successful use.

Introduction

Soil erosion represents the removal of soil from the location where it developed by erosion processes. These processes are driven by several agents, the most important of which are water and wind. Water erosion is the most widespread cause of soil loss and occurs in a wide range of environments. Wind erosion can be equally destructive, but its occurrence is more limited. Soil erosion can be viewed as a natural process, but rates of soil loss are commonly very low in natural undisturbed landscapes, where the soil will generally be well protected by vegetation from the destructive effects of rainsplash, surface runoff, and deflation. Under these conditions, the rate of soil loss is likely to be balanced by the rate of soil formation and the soil will be in a quasi-equilibrium. When, however, the vegetation cover is removed or reduced by human activities, such as land clearance and cultivation, excessive grazing or forestry activity, rates of soil loss can increase by several orders of magnitude. Such accelerated soil erosion can represent a serious threat to the sustainability of the soil resource and to food security, through both reduced crop productivity and the loss of agricultural land as a result of soil removal, gully development, and other destructive effects of erosion. Recent estimates indicate that ca. 1,000 Mha of the earth's surface is currently adversely affected by water erosion and ca. 500 Mha by wind erosion, with about 80% of these totals being associated with the developing world [1]. Much of this erosion occurs on agricultural land. It has been estimated that about 80% of the world's agricultural land suffers from moderate-to-severe erosion and that about 10% suffers from slight-to-moderate erosion [2]. Even in developed countries, such as the USA, where soil conservation measures are widely applied, soil erosion represents a major problem, and it has been reported that in the USA, about 23 million hectares of cropland experience excessive soil erosion and that erosion rates exceed the tolerable soil loss on about a further 20 million hectares [3].

The annual global loss of agricultural land due to accelerated erosion has been estimated to be ca. 3 million hectares [4] and the progressive destruction of the global soil resource by erosion is emphasized by the calculation that the world's croplands are losing about 23 billion tons of soil in excess of new soil

formation each year [5]. The rate of soil formation will be influenced by a wide range of factors, but a value of $1 \text{ t ha}^{-1} \text{ year}^{-1}$ is frequently cited as a typical value. This value is often referred to as the soil loss tolerance, since, if erosion rates exceed this value, the soil depth will progressively reduce and crop productivity is likely to decline. Global models have been used to assess the extent to which this value of $1 \text{ t ha}^{-1} \text{ year}^{-1}$ is exceeded and, based on the global pattern of land use in the 1980s, the global mean soil erosion has been estimated to be $10.2 \text{ t ha}^{-1} \text{ year}^{-1}$, a value which is an order of magnitude greater [6]. Average erosion rates were shown to be highest in Asia ($12.2 \text{ t ha}^{-1} \text{ year}^{-1}$) and lowest in Australia ($3.0 \text{ t ha}^{-1} \text{ year}^{-1}$).

Current concern for the impact of climate change and global warming has directed attention to likely changes in soil erosion rates. In most areas of the world, predictions suggest that soil erosion rates will increase and, in the global modeling exercise referred to above [6], climate change predictions associated with a doubling of CO_2 concentrations were coupled with erosion models to estimate that at the global scale, soil loss would increase by 9% by the 2090s as a result of climate change. When this prediction was coupled with likely changes in land use by the 2090s, soil erosion increased by 14%. Soils play a key role in the global carbon cycle by storing carbon. Accelerated erosion perturbs this cycle by mobilizing soil carbon that may be subsequently oxidized causing increased CO_2 emissions, and by redistributing soil carbon and sequestering it in longer-term depositional sinks [7, 8]. Both past and likely future increases in soil erosion therefore have important implications for the global carbon cycle.

Although concern for soil erosion has traditionally focused on the on-site effects linked to loss of agricultural land and reduced productivity, it is important to recognize that much of the sediment mobilized by accelerated water erosion will find its way into streams and river systems leading to increased sediment loads. Increased sediment loads can give rise to a range of environmental problems linked to degradation of water quality and aquatic habitats, damage caused to infrastructure during floods, and increased deposition of sediment in river channels, lakes, and reservoirs [9]. These impacts are frequently referred to as off-site impacts, to distinguish them from the on-site effects

highlighted above. In many situations, the off-site impacts of soil erosion will be more serious than the on-site effects. Taken together, they represent a very large economic cost and it has been estimated that worldwide, the annual cost of both the on-site and off-site impacts of soil erosion on agricultural land amounts to about \$400 billion [10]. However, it is difficult to quantify the economic cost of loss of biodiversity, changes in the carbon cycle, and impacts on ecosystem services, and the real cost may therefore be substantially greater.

The deposition of sediment within the landscape and in water bodies, such as rivers, lakes, reservoirs, and marshes, must be seen as an inevitable complement to the mobilization of sediment by erosion and such processes are frequently referred to as sedimentation. Most attention is generally directed to sedimentation associated with water erosion, since sediment mobilized by such erosion can be deposited before reaching the stream network and, once entering the river system, there will be many further opportunities for sedimentation downstream. A substantial proportion of the sediment moving through the river system will reach the outlet of the river basin and may then be deposited in estuaries, in coastal areas and in the deeper ocean. In contrast, soil mobilized by wind erosion is usually deposited locally, although the finer fractions can be transported over substantial distances and it may be less easy to link erosion and sedimentation. Equally, such sediment is likely to be deposited over large areas and its impact is likely to be much less. Again, sedimentation must be seen as a natural process, but accelerated erosion will frequently be coupled with increased sedimentation, which can lead to environmental problems. These problems relate not only to the physical presence of deposited sediment, which can, for example, greatly reduce the water holding or storage capacity of a water body, but also to the nutrients and contaminants, such as pesticides, associated with the sediment. Release of nutrients and contaminants from sediment deposited in water bodies can cause eutrophication and pollution of those water bodies. Similarly, deposition of contaminated sediment or sediment with a high-nutrient content on river floodplains can cause contamination of soil and crops and changes in the species composition of native vegetation communities, in response to the changing nutrient status of

the soil. Estimates suggest that the world's large reservoirs are currently losing about 1% of their total storage each year [11]. Such losses can have important implications for water supply and the annual cost of replacing this lost storage has been estimated to be ca. \$6 billion. When this cost is combined with the problems of finding new dam sites, when existing reservoirs are filled with sediment and can no longer operate effectively, the serious consequences of reservoir sedimentation are clear.

Against this background, there is an increasing need to assemble information on soil erosion and sedimentation rates to inform and support the sustainable management of soils, catchments, and associated water bodies. Such demands have increased with the development and use of numerical models as management tools, since such models commonly require data for calibration and validation. Although a range of techniques have been developed for documenting soil erosion and sedimentation rates [12, 13], these techniques have important limitations. For example, in the case of soil erosion, erosion plots are widely used to obtain information on rates of soil loss under particular conditions. (e.g., different soil types, slope gradients, crop types, cultivation practices, etc.). However, such plots are costly to install, they need to be operated for an extended period to obtain reliable information on mean annual rates of soil loss due to the inherent temporal variability of erosional events, and the results obtained from a small bounded plot may be unrepresentative of the natural landscape with its more complex topographic form. In the latter context, plot measurements relate to the export of material across the lower boundary of the plot and thus only provide spatially averaged estimates of net soil loss. They cannot provide information on deposition of mobilized soil within the plot and thus soil redistribution rates more generally (i.e., gross erosion, deposition, and net erosion). Furthermore, soil redistribution rates may vary markedly according to location within a field in relation to the local topography (e.g., ridges and hollows and concave and convex slope profiles). To meet the requirements of the current generation of distributed numerical models, spatially distributed information on erosion and deposition rates will frequently be required. Similar problems are faced by attempts to document deposition rates on river

floodplains and in water bodies, such as lakes, reservoirs, and wetlands. Short-term variation of sediment deposition rates and spatial variability of deposition rates may introduce the need to quantify such variability.

The limitations of traditional measurement techniques referred to above have directed attention to the potential for using environmental radionuclides to provide information on soil erosion and sedimentation rates. The fallout radionuclides cesium-137 (^{137}Cs), unsupported or excess lead-210 ($^{210}\text{Pb}_{\text{ex}}$), and beryllium-7 (^7Be) have proved particularly useful in this context, because they are rapidly and strongly fixed on reaching the soil or sediment surface as fallout and their post-fallout redistribution provides a means of documenting soil and sediment redistribution within the landscape. In essence, measurements of the radionuclide areal activity density or inventory (Bq m^{-2}) are made at a number of points across a field by collecting soil cores. By comparing the results of these measurements with a measurement of the reference inventory, which represents the inventory found in an adjacent undisturbed area, areas or points experiencing erosion and deposition can be identified. In the former case, the inventory at the sampling point will be less than the reference inventory, and in the latter case, the inventory will exceed the reference inventory. The degree to which the inventory departs from the reference inventory is a direct reflection of the magnitude of the erosion or deposition rate, and a range of conversion models have been developed to estimate erosion and deposition rates from the measurements obtained. In a similar way, fallout inputs to water bodies, which are subsequently incorporated into sediment deposits can provide a means of establishing the chronology of sediment deposits. In the case of ^{137}Cs , the occurrence of peak fallout in 1963 provides a time marker for that year and the rate of decrease of $^{210}\text{Pb}_{\text{ex}}$ activity with depth affords a means of estimating the time elapsed since the deposition of sediment from a specific depth and thus the sedimentation rate.

Early work on the use of fallout radionuclides to document erosion and sedimentation processes focused on ^{137}Cs and dates from the aftermath of the bomb testing of the 1950s and early 1960s. Cesium-137 can provide information relating to the redistribution of fallout received in the late 1950s and the 1960s,

which currently provides a timeframe of ca. 50 years. The use of ^{137}Cs has been complemented more recently by the use of $^{210}\text{Pb}_{\text{ex}}$ and ^7Be , which provide information relating to longer (i.e., ca. 100 years) and much shorter (i.e., the past few weeks) timeframes, respectively [14]. Fallout radionuclides are being increasingly used in soil erosion and sedimentation investigations in many different areas of the world, and they are able to provide essentially unique information on rates and spatial patterns of erosion and sedimentation that cannot be obtained using traditional approaches. Particular advantages in terms of documenting erosion and sedimentation and particularly soil erosion include the ability to:

- Provide retrospective information on average rates of soil redistribution or sediment deposition (e.g., for the past 50 or 100 years) on the basis of contemporary sampling often involving only a single site visit.
- Provide both time-integrated data relating to the past 50 or 100 years (^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$) and data relating to individual events or short periods of heavy rainfall occurring at present (^7Be).
- Integrate the various processes contributing to the redistribution of soil and sediment particles in the landscape.
- Obtain information from the “natural” landscape without the need for special installations or significant disturbance of the study area.
- Avoid the need for costly long-term monitoring.
- Assemble spatially distributed information that is directly compatible with recent advances in the development of physically based distributed models and the application of GIS and geostatistics to modeling erosion and sedimentation.
- Provide an approach that is applicable to several different components of the system, including erosion, sediment delivery, and sediment storage and deposition, as well as to tracing sediment sources. In the case of water erosion, it is also able to integrate processes associated with sheet erosion, rill erosion, and tillage erosion.
- Provide an approach that is applicable throughout the world, since fallout radionuclide fluxes and inventories are of sufficient magnitude to permit their use in most areas of the globe.

This entry provides an overview of the basis for using fallout radionuclides in the study of erosion and sedimentation and of the information that can be obtained. It provides a brief history of the approach; a description of the key characteristics of the individual fallout radionuclides, including their source, fallout input, and post-fallout behavior; a discussion of their use to document soil erosion and soil redistribution, as well as sediment deposition; a brief review of field and laboratory techniques and key issues that need to be addressed; and some examples of applications, based on the work of the author over the past 30 years.

A Brief History of the Use of Fallout Radionuclides in Erosion and Sedimentation Studies

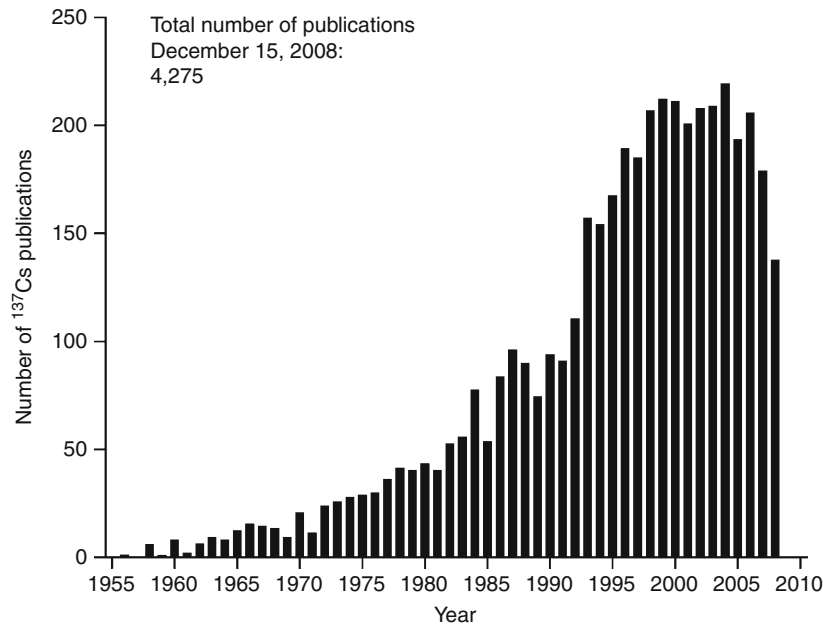
As indicated above, work on the use of fallout radionuclides in soil erosion and sedimentation investigations largely dates from the aftermath of the nuclear weapons testing that occurred in the 1950s and early 1960s, although early work on the use of $^{210}\text{Pb}_{\text{ex}}$ to establish the chronology of sediment cores [15] developed essentially independently of this. The nuclear weapons testing that occurred in the 1950s and 1960s caused the fallout of a range of radionuclides, including strontium-90 (^{90}Sr) and ^{137}Cs . Menzel was arguably the first to recognize the link between radionuclide loss and soil loss from plots [16], although his work focused on ^{90}Sr . Subsequent studies [17–19] highlighted the link between ^{137}Cs loss and soil loss, and McHenry and Ritchie demonstrated that the spatial pattern of ^{137}Cs inventories within a field could be used to identify areas of erosion and deposition [20]. Based on this early work, the potential for using ^{137}Cs to document soil erosion and sediment redistribution rates has been successfully exploited in many different areas of the world. In addition to the very important lead provided by Ritchie working in the USA, particular reference can be made to pioneering work undertaken in Canada [21, 22], Australia [23–25], and the UK [26, 27]. The Chernobyl accident in 1986 also provided a significant input to the expansion of work on the application of ^{137}Cs measurements, by directing renewed attention to the behavior and fate of radiocesium in the environment and providing new opportunities to exploit the approach [28]. A further important stimulus to work in the field was provided by the International Atomic

Energy Agency (IAEA). This body sponsored two Coordinated Research Projects between 1995 and 2001 aimed at promoting and standardizing procedures for using ^{137}Cs measurements in soil erosion and sedimentation investigations. These two Coordinated Research Projects culminated in the publication of a handbook for the assessment of soil erosion and sedimentation using environmental radionuclides that focused on the use of ^{137}Cs [29].

A clear and useful demonstration of the near exponential growth of studies involving the application of ^{137}Cs in soil erosion and sedimentation investigations between the late 1960s and the mid 1990s is provided by the number of entries for specific years included in the bibliography of publications of ^{137}Cs studies related to erosion and sediment deposition, compiled by Ritchie and Ritchie [30]. Although some of the publications included in this bibliography relate to broader aspects of the behavior of ^{137}Cs in the environment and to related radionuclides, the trend demonstrated in Fig. 1 provides a graphic indication of the rapid expansion of work in this field.

Although they focused attention on ^{137}Cs , the two IAEA Coordinated Research Projects referred to above also aimed to promote the use of other fallout radionuclides, and more particularly ^7Be and $^{210}\text{Pb}_{\text{ex}}$, in soil erosion investigations. The development of studies using these two fallout radionuclides commenced in the early and mid-1990s and important early work includes that of Wallbrink and Murray [31], Blake et al. [32], Walling et al. [33], Matisoff et al. [34], and Wilson et al. [35] with ^7Be , and Wallbrink and Murray [36], Walling and coworkers [37–39] and Matisoff et al. [34] with $^{210}\text{Pb}_{\text{ex}}$. Important drivers in promoting this work with additional fallout radionuclides included the possibility of addressing different timescales, with $^{210}\text{Pb}_{\text{ex}}$ covering up to ca. 100 years and ^7Be providing information on individual events and short periods of heavy rainfall. Another driver was the opportunity to increase the potential for using fallout radionuclides in areas of the world where ^{137}Cs inventories are very low and where these inventories will decline further in the near future, due to radioactive decay and the lack of contemporary fallout inputs.

Application of ^{137}Cs measurements to the study of sedimentation rates in depositional environments and to establishing the recent chronology for sediment



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 1

The growth in the annual number of publications reporting work involving ^{137}Cs listed in the bibliography produced by Ritchie and Ritchie [30]

cores from lakes and reservoirs can again be traced back to the immediate aftermath of the bomb tests of the late 1950s and early 1960s [40]. The potential for using the peak fallout input in 1963 (or 1964 in the southern hemisphere) as a time marker was rapidly exploited by many investigators, and the approach was applied to a range of depositional sinks, in addition to lakes and reservoirs, including wetlands [41, 42], estuaries and coastal saltmarshes [43, 44], and river floodplains [45, 46]. The further fallout input provided by the Chernobyl accident introduced a second time marker in many areas of Europe and adjacent areas, which has improved the temporal resolution that can be achieved. Work by limnologists and sedimentologists on the application of $^{210}\text{Pb}_{\text{ex}}$ in establishing the chronology of sediment cores over the past ca. 100 years again developed in the early 1970s [47, 48] but proceeded essentially independently of work with ^{137}Cs . This in part reflects the different principle involved, namely, the reduction in activity with depth in response to decay, which was not dependent on the timing of weapons testing. The first application of the approach was to the accumulation of ice sheets and was described by Goldberg in the early 1960s [15]. Since the 1970s,

key developments have included the use of direct measurements of $^{210}\text{Pb}_{\text{ex}}$ using gamma spectrometry as an alternative to its measurement using alpha spectrometry, which is considerably more demanding in terms of the chemical extraction required, and the development of improved models for interpreting the $^{210}\text{Pb}_{\text{ex}}$ depth distribution and establishing the chronology [49–51]. Because of its short half-life (53 days), ^7Be is commonly limited to the thin upper layer of sediment cores and is of little value for dating or establishing sedimentation rates. However, this radionuclide has been used since the 1980s to distinguish very recent sediment from older sediment and to quantify sediment mixing [52] and to estimate depths of overbank sedimentation on river floodplains associated with individual flood events [53].

Fallout Sources and Inputs

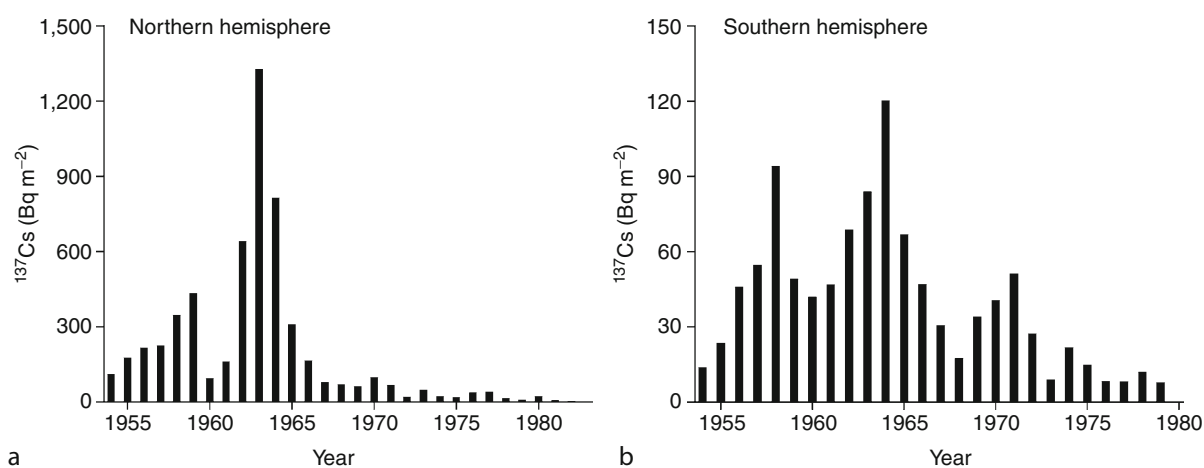
As indicated above, this contribution focuses on the application of three fallout radionuclides, namely, ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$ and ^7Be , in soil erosion and sedimentation investigations. These three radionuclides differ in terms of their half-lives, their sources, their fallout inputs, and the temporal variation of the resulting soil

inventories, and it is these differences that mean that they are able to be used to provide different, but complementary, information. It is important that their sources and the key features of their fallout input should be understood, since they have important implications for the interpretation of the measurements used to derive estimates of erosion and sedimentation rates. A brief overview of the source and fallout input of the three radionuclides is therefore provided below.

Cesium-137

Cesium-137 ($t_{1/2}$ 30.1 years) differs from the other two radionuclides in that whereas they are of natural origin, this radionuclide is manmade. It is found in the global environment primarily as a result of the testing of thermonuclear weapons (bomb tests) that took place during the period extending from the early 1950s to the early 1960s. In addition, a further fallout input occurred in parts of Europe and adjacent regions as a result of the Chernobyl accident in 1986. The bomb tests injected ^{137}Cs into the stratosphere, where it mixed and circulated globally before being redeposited as fallout. The first high-yield bomb test took place in 1952, but ^{137}Cs fallout was below detection in most areas of the world prior to 1954. The subsequent temporal pattern of fallout reflected the magnitude and number of bomb tests and the residence time of the

^{137}Cs in the stratosphere (ca. 1–10 years). Figure 2 summarizes available data on the temporal distribution of the annual fallout resulting from the bomb tests in the 1950s and 1960s, by combining data for New York, USA, and Milford Haven, UK to provide a typical record for the northern hemisphere, and from Adelaide and Brisbane to provide a typical record for the southern hemisphere. ^{137}Cs was not routinely monitored during the main period of bomb fallout, but reliable estimates can be derived from the measurements of ^{90}Sr fallout made at this time, since the two radionuclides are found in a fixed ratio. Most bomb tests were located in the northern hemisphere and Fig. 2 shows a gradual increase in ^{137}Cs fallout in the northern hemisphere up to 1959, when the temporary moratorium on bomb tests imposed in 1958 caused a reduction in fallout, followed by a further increase in fallout up to 1963 when the Nuclear Test Ban Treaty resulted in a cessation of bomb tests and fallout began to decline. Fallout levels fell below detection by the early 1980s. The further fallout input that occurred in 1986, as a result of the Chernobyl accident, was localized and more short-lived, because the radiocesium plume released from the accident was only dispersed into the troposphere. The data presented in Fig. 2 are estimates of annual fallout input. During a given year, the pattern of fallout is likely to have reflected that of precipitation, because most of the fallout occurred as wet fallout.



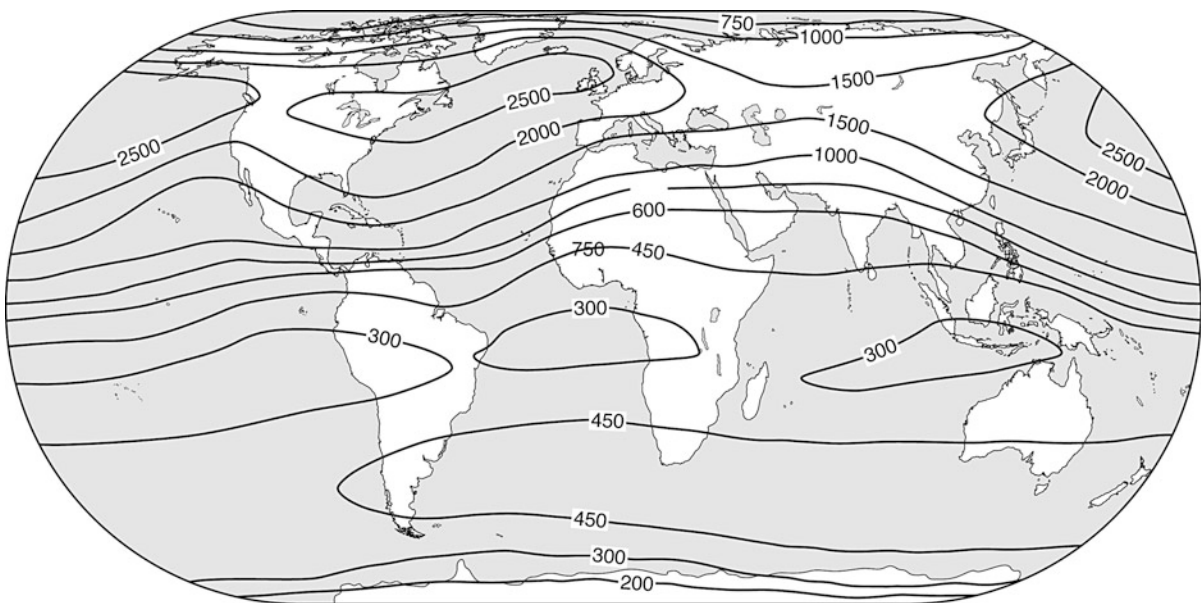
Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 2

Typical ^{137}Cs fallout records for the northern (New York, USA/Milford Haven, UK) and southern (Adelaide/Brisbane, Australia) hemispheres

The bomb tests injected ^{137}Cs into the stratosphere, and it was globally distributed. The global pattern of fallout reflects the global pattern of precipitation, since most fallout occurs as wet fallout in association with precipitation, the location of the bomb tests, and the pattern of stratospheric mixing. Because most tests occurred in the northern hemisphere, the global pattern of fallout evidences a clear latitudinal zonation. **Figure 3** presents a generalized map of the global pattern of ^{137}Cs fallout associated with the bomb tests, as represented by the inventories found in 2000. Maximum fallout occurred in the mid-latitudes of the northern hemisphere and minimum fallout occurred around the equator. Overall, the amount of fallout received in the southern hemisphere was only about 30% of that received in the northern hemisphere. This feature of the global fallout pattern introduces some limitations on use of ^{137}Cs in the southern hemisphere, since the considerably lower inventories can introduce some measurement constraints.

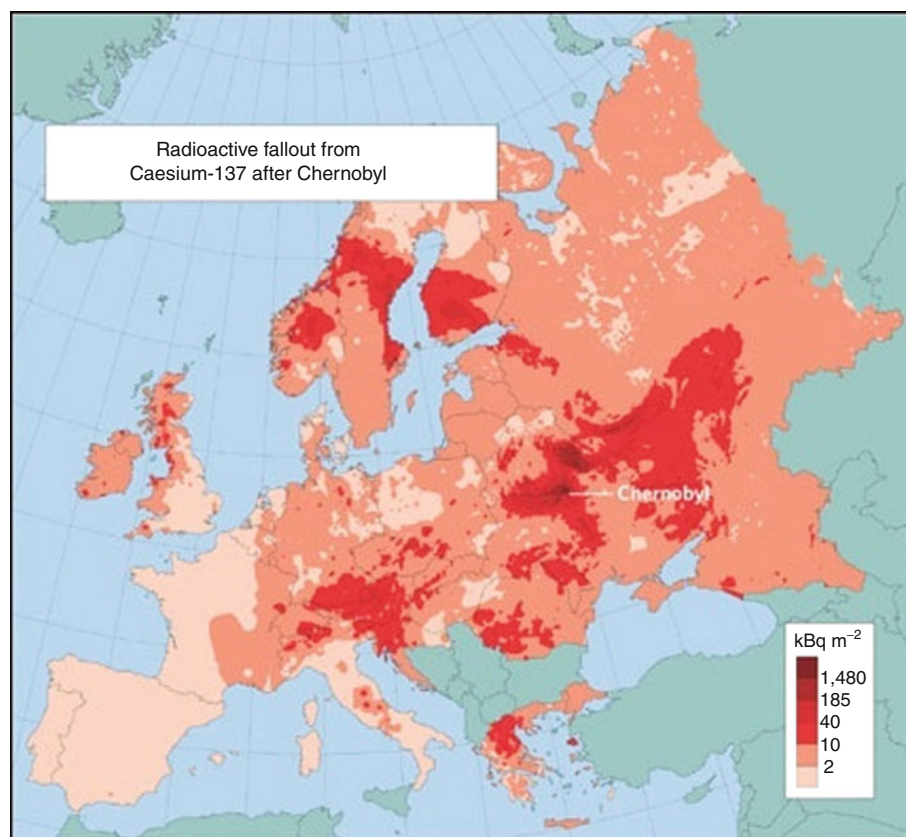
It has been estimated that the amount of ^{137}Cs released by the Chernobyl disaster was of the order of 15% of that produced by the bomb tests. Because of its more limited tropospheric transport and dispersal,

significant Chernobyl fallout is found over a more limited area of the globe located around the Chernobyl source (see **Fig. 4**). Here the pattern reflects the trajectory of the plume of contaminant released from Chernobyl and the distribution of precipitation during the immediate post-accident period. Considerable local variability has been reported in the case of Chernobyl fallout and this partly reflects the fact that the fallout was associated with a small number of rainfall events occurring within a short period. This contrasts with the situation for bomb fallout, which was distributed over a ca. 20-year period during which local variability is likely to have been smoothed. Reports of the magnitude of the depositional fluxes that occurred in the immediate vicinity of Chernobyl suggest that these were as high as $1,480 \text{ kBq m}^{-2}$ and therefore more than two orders of magnitude greater than the total fallout received in areas of high bomb fallout. In many areas affected by Chernobyl fallout, that fallout input will now dominate the total inventory and this must be taken into account when using ^{137}Cs measurements to document soil erosion and sedimentation rates. However, there are also many areas in Europe where the two fallout inputs are of a similar magnitude and this may



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 3

The global pattern of ^{137}Cs inventories (Bq m^{-2}) in 2000 (based on Walling and He [54])



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 4

The pattern of ^{137}Cs fallout over Europe resulting from the Chernobyl disaster (Based on Smith and Beresford [55])

introduce problems when interpreting contemporary measurements of ^{137}Cs inventories. In the period immediately after the Chernobyl accident, it was possible to distinguish the two components of the total ^{137}Cs inventory by also measuring the cesium-134 (^{134}Cs) activity. ^{134}Cs ($t_{1/2}$ 2.06 years) occurred in a fixed ratio to ^{137}Cs in Chernobyl fallout and this provided a means of quantifying the proportion of the total ^{137}Cs inventory attributable to Chernobyl fallout. Because of its short half-life, this approach could only be used for a few years immediately after the Chernobyl accident. However, scope exists to use the plutonium ($^{239,240}\text{Pu}$) content of the current inventory to establish the relative contribution of Chernobyl and bomb fallout, since $^{239,240}\text{Pu}$ was absent from Chernobyl fallout.

As emphasized by Fig. 2, a key feature of ^{137}Cs fallout is that in most areas of the globe, virtually

all of this fallout input occurred during the period 1955–1970, with a peak in 1963. The basis for using this radionuclide to obtain estimates of erosion and sedimentation rates reflects this pattern of input. By studying the subsequent redistribution of the ^{137}Cs fallout in the landscape, it is possible to obtain information on soil and sediment redistribution over the past ca. 45 years. In addition, the occurrence of peak fallout in 1963 provides a valuable time marker in many depositional environments. Additional fallout inputs associated with the Chernobyl accident may mean that the main fallout input occurred in 1986 or if the inputs were of a similar magnitude that there were two periods of fallout input, one extending from 1955 to 1970 and the other occurring in 1986. The occurrence of the second fallout input can introduce a second time marker in sediment deposits which can prove very useful in establishing the chronology of a core.

Excess Lead-210

Lead-210 (^{210}Pb , $t_{1/2}$ 22.3 years) is a natural geogenic radionuclide that is produced in the Uranium-238 (^{238}U) decay chain. Radium-226 (^{226}Ra , $t_{1/2}$ 1,600 years) is a product of that chain and is found in most soils and rocks. This decays to gaseous Radon-222 (^{222}Rn , $t_{1/2}$ 3.8 days) which in turn undergoes a series of short-lived decays to ^{210}Pb . During its short half-life, some of the ^{222}Rn escapes from the soil into the overlying atmosphere, where it decays to ^{210}Pb , which returns to the earth's surface as fallout. The residence time of ^{210}Pb in the atmosphere is estimated to be only a few days. The ^{222}Rn that remains in the soil also decays to ^{210}Pb and, since this is an in situ product of the ^{226}Ra , it is commonly referred to as supported ^{210}Pb . It is possible to calculate the amount of supported ^{210}Pb that should theoretically exist in the soil, given the level of ^{226}Ra , by assuming equilibrium between the two radionuclides. However, the actual amount is likely to be less, since, under most conditions, some of the ^{222}Rn would have escaped to the overlying atmosphere. Without such escape or exhalation, supply of ^{222}Rn to the atmosphere and subsequent ^{210}Pb fallout would not occur. To distinguish the ^{210}Pb reaching the soil surface as fallout from that produced in situ within the soil, the former is commonly referred to as unsupported or excess ^{210}Pb ($^{210}\text{Pb}_{\text{ex}}$). The total inventory (Bq m^{-2}) of ^{210}Pb contained in a given depth of soil can therefore be partitioned into a supported and an unsupported or excess component. The former can be estimated from the amount of ^{210}Pb that would be expected if it was in equilibrium with the ^{226}Ra , reduced by a factor reflecting the proportion of the ^{222}Rn that escaped to the overlying atmosphere. It is not easy to establish this factor empirically, but it can be estimated by considering the difference between the theoretical estimate of supported ^{210}Pb activity (assuming equilibrium between ^{226}Ra and ^{210}Pb) in soil or sediment below the depth influenced by surface inputs of $^{210}\text{Pb}_{\text{ex}}$ and the measured activity at that depth. Since the fallout radionuclide (i.e., $^{210}\text{Pb}_{\text{ex}}$) is the component of interest for use in documenting erosion and sedimentation rates, it is important to direct attention to this, rather than the total ^{210}Pb activity. Other sources of ^{210}Pb fallout include exhalation of ^{222}Rn from the surface of the oceans, release of ^{226}Ra ,

^{222}Rn , and ^{210}Pb from volcanic eruptions, and industrial emissions. Because the amount of ^{210}Pb produced over the oceans as a result of release and decay of ^{222}Rn is about two orders of magnitude less than that over land, this source is of limited importance for terrestrial fallout. The contribution of volcanic eruptions and industrial emissions to global fallout of ^{210}Pb is again minimal, but these sources can be locally significant.

In contrast to ^{137}Cs , the annual fallout of $^{210}\text{Pb}_{\text{ex}}$ at a given location can be viewed as being essentially constant in the medium term. If it is assumed that the total fallout of $^{210}\text{Pb}_{\text{ex}}$ is retained in the upper horizons of the soil of a stable reference site, that experiences neither erosion nor deposition, the mean annual fallout input can be estimated by assuming a steady-state equilibrium between fallout input and loss through radioactive decay. There are few longer-term records of annual $^{210}\text{Pb}_{\text{ex}}$ fallout involving direct measurements that can be used to assess the degree of inter-annual variability, but a record from Munich-Neuherberg in southern Germany for the period 1981–1999 [56] demonstrated a twofold variation in annual fallout, with values ranging from 120 to 250 $\text{Bq m}^{-2} \text{ year}^{-1}$, with a mean of 180 $\text{Bq m}^{-2} \text{ year}^{-1}$. A similar factor of two for the inter-annual variability of ^{210}Pb fallout has been reported by other studies. This inter-annual variability can be partly accounted for by variations in annual precipitation, but the relationship between the annual ^{210}Pb fallout flux and annual rainfall is not always well defined. The distribution of fallout within the year will reflect both the precipitation regime and seasonal variation of the concentration of $^{210}\text{Pb}_{\text{ex}}$ in the atmosphere, as well as the degree of mixing of the atmosphere and the volume of air scavenged by the rainfall. High concentrations are often found in the atmosphere during late spring and summer when drier soils and the absence of snow cover increase the exhalation of ^{222}Rn to the atmosphere from the soil [57]. Valles et al. [58] report a twofold variation in monthly ^{210}Pb concentrations in air sampled at Barcelona, Spain and at this location, concentrations were substantially higher in summer and autumn. However, equivalent information on air concentrations reported for Munich-Neuherberg [56] showed a contrasting pattern, with maximum values between September and February. At this site, maximum

monthly fallout fluxes occurred in the summer months between June and August, in response to the higher precipitation during these months.

Less is currently known about the global variation of $^{210}\text{Pb}_{\text{ex}}$ fallout than for ^{137}Cs . Values of annual fallout reported in the literature range over more than an order of magnitude from $<50 \text{ Bq m}^{-2} \text{ year}^{-1}$ to in excess of $500 \text{ Bq m}^{-2} \text{ year}^{-1}$ (see Table 1). In terms of the global pattern, available information suggests that, although annual fallout fluxes will reflect annual rainfall, they are generally lower in the western areas of the continents, where the dominant air mass movements are off the oceans where very little ^{222}Rn will be produced. In contrast, higher annual fluxes are commonly found in the eastern areas of continents where the air mass tracks pass over large areas of land where ^{222}Rn is produced and concentrations of dust increase. A clear latitudinal zonation in annual fallout of ^{210}Pb has also been reported [59], with much lower values being documented in the southern hemisphere, due to the greater expanse of ocean. The values listed in Table 1 are consistent with these general controls. However, it is important to note that there may be appreciable local variability in annual fallout deposition. Values of annual ^{210}Pb deposition of 59 and $85 \text{ Bq m}^{-2} \text{ year}^{-1}$ in 1990 have been reported for two sites in the southern Netherlands that are only 5 km apart [60].

Beryllium-7

Beryllium-7 (^7Be $t_{1/2}$ 53.3 days) is, like ^{210}Pb , a naturally occurring fallout radionuclide. In this case, its origin is cosmogenic and it is produced in the troposphere and stratosphere by the cosmic ray spallation of nitrogen and oxygen. The production of ^7Be is controlled by the cosmic ray flux which varies with altitude, latitude, and solar activity [66]. Production is reported to be greatest at altitudes between 12 and 20 km, depending on latitude, and decreases almost exponentially toward the earth's surface, where it is about three orders of magnitude lower [67]. Globally, production increases from the equator toward the poles and this increase is characterized by a factor of approximately 1–3 for the troposphere and 4–5 for the stratosphere [67]. Because ^7Be production is influenced by solar activity, it has been linked to sunspot numbers and the 11-year solar cycle. The increased incidence of sunspots reduces

Fallout Radionuclides and the Study of Erosion and Sedimentation. Table 1 Mean annual deposition of $^{210}\text{Pb}_{\text{ex}}$ at different locations

Location	Mean ^{210}Pb fallout ($\text{Bq m}^{-2} \text{ year}^{-1}$)
Tatsunokuchi, Japan ^a	840
Hokkaido, Japan	367
Bombay, India	250
Munich, Germany ^b	180
Galveston, USA ^c	172
New Haven, USA	153
Geneva, Switzerland ^d	150
Delhi, India	133
Moscow, Russia	115
Bermuda ^e	115
Calcutta, India	102
Darwin, Australia	95
Milford Haven, UK	85
Suva, Fiji	80
Groningen, the Netherlands ^f	71
Alice Springs, Australia	57
Sydney, Australia	53
Auckland, New Zealand	50
East Antarctica ^g	31

Adapted from Beks et al. [60], with additions.

^aFrom Turekian et al. [61]

^bFrom Winkler and Rosner [56]

^cFrom Yamamoto et al. [62]

^dFrom Baskaran et al. [63]

^eFrom Beks et al. [60]

^fFrom Preiss et al. [59]

^gFrom Caillet et al. [64]

the cosmic ray flux to the earth and thus ^7Be production. Because of the longer residence times of aerosols in the stratosphere, ^7Be activity remains relatively constant in the stratosphere. However, with lower production rates and shorter aerosol residence times due to rapid washout, ^7Be concentrations in the troposphere are one or two orders of magnitude less than that in the stratosphere. Available evidence suggests that >90% of ^7Be deposition occurs as wet deposition [68].

There are relatively few records of ^7Be fallout to provide a detailed basis for defining its characteristics, but available evidence indicates that in addition to inter-annual variability connected to sunspot activity, ^7Be concentrations in precipitation and fallout fluxes vary seasonally. These variations will reflect the incidence of mixing between the stratosphere and troposphere, the altitude of precipitation generation, and precipitation amounts. Maximum concentrations and fluxes are often recorded in spring and early summer in mid-latitudes, in response to increased stratosphere-troposphere exchange and increased convection due to warming of the land surface. Considerable variability in ^7Be concentrations may be associated with individual rainfall events, with some researchers reporting that activities can vary by a factor of greater than 20 between events. However, fairly well-defined relationships between ^7Be depositional flux and event rainfall are

commonly found [64]. Because of its short half-life, it is more important to understand the magnitude and timing of fallout inputs, since these will have a much greater effect on the soil inventory than in the case of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$, where the inventory of these radionuclides will remain essentially constant in the short term and reflect the long-term pattern of fallout.

As with ^{210}Pb , there are currently insufficient measurements to obtain a clear picture of the global pattern of annual ^7Be deposition. However, available measurements indicate that annual fluxes range between ca. 500 and 6,500 $\text{Bq m}^{-2} \text{ year}^{-1}$ (see Table 2). The influence of annual rainfall amount is readily demonstrated by the high annual deposition of 6,350 $\text{Bq m}^{-2} \text{ year}^{-1}$ reported for Hokitika in South Island, New Zealand [68], where the annual rainfall was ca. 2,800 mm. Since most available measurements are from the mid-latitudes, it is difficult to identify the latitudinal

Fallout Radionuclides and the Study of Erosion and Sedimentation. Table 2 Annual deposition of ^7Be recorded at different locations

Place	Mean annual precipitation (mm)	Annual ^7Be fallout ($\text{Bq m}^{-2} \text{ year}^{-1}$)	Source
Hokitika, New Zealand	2,800	6,350	Harvey and Matthews [69]
Tatsunokuchi, Japan	–	5,300	Yamamoto et al. [62]
New Haven, CT, USA	1,390	3,780	Turekian et al. [70]
Bermuda	1,700	2,850	Turekian et al. [70]
Galveston, USA	1,167	2,450	Baskaran et al. [63]
Geneva, Switzerland	966	2,087	Caillet et al. [64]
Norfolk, VA, USA	1,313	2,075	Todd et al. [71]
Milford Haven, UK	1,328	1,618	Peirson [72]
Bombay, India	2,277	1,267	Lal et al. [73]
Heidelberg, Germany	810	1,249	Schumann and Stoeppler [74]
Canberra, South Eastern Australia	660	1,030	Wallbrink and Murray [68]
Chilton, UK	822	898	Peirson [72]
Arkansas, USA	1,070	867	Lee et al. [75]
Westwood, NJ, USA	787	717	Walton and Fried [76]
Damascus, Syria	153	528	Othman et al. [77]
East Antarctica	–	700	Nijampurkar and Rao [65]
Malaga, Spain	308	412	Duenas et al. [78]

influence that could be expected to result from the latitudinal variation in ^7Be production noted above. Because of its cosmogenic origin and the importance of production in the stratosphere, ^7Be deposition fluxes can be expected to be similar in both the northern and southern hemispheres. This contrasts with ^{137}Cs and ^{210}Pb , where the location of the bomb tests and the much larger extent of the oceans in the southern hemisphere, respectively, exert an important influence in reducing fallout inputs in the southern hemisphere.

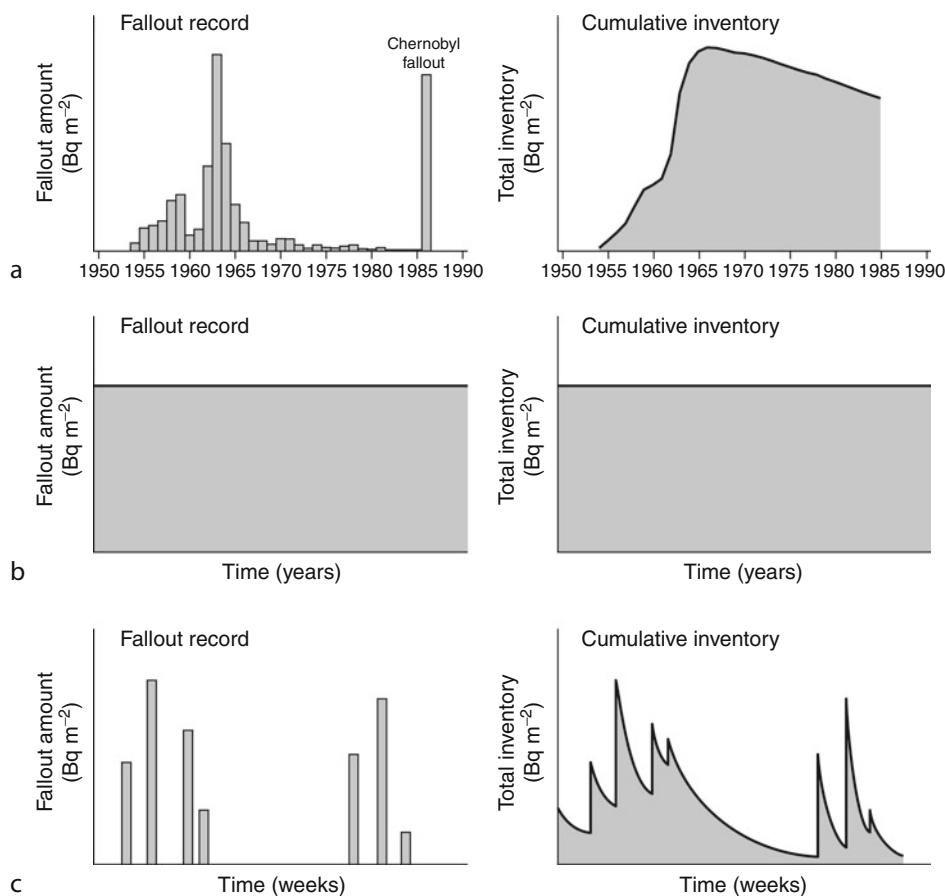
Fallout Radionuclides in the Soil and Soil Inventories

^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$, and ^7Be fallout reaches the soil surface as positively charged cations, and in most environments, the fallout will be rapidly and strongly adsorbed by mineral particles and organic material and become essentially non-exchangeable [79–83]. Preferential adsorption by clay minerals and fine particles has been reported in many studies [84]. The strong affinity of these radionuclides for the soil is reflected by their high partition coefficients (K_d) which are commonly of the order of 10^4 – 10^5 L kg^{-1} . Existing evidence suggests that whereas ^{137}Cs is primarily associated with the mineral fraction of soils, the organic fraction is also important in the uptake of ^7Be and $^{210}\text{Pb}_{\text{ex}}$. The speed of uptake and the kinetics of the adsorption are less well defined, but time spans for uptake or adsorption ranging from a few minutes to less than an hour are often cited. It is, however, important to recognize that in some environments, the fallout radionuclides reaching the soil surface may be less strongly fixed. This, for example, appears to be case for ^{137}Cs and peat soils, because there are a number of reports of Chernobyl fallout being poorly fixed by peat soils and remaining mobile within the soil and runoff.

Where the soil supports a vegetation canopy, this will complicate the transfer of the radionuclides to the surface soil. In the case of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ where emphasis is usually placed on the longer-term accumulation and fate of the fallout input in the landscape, such interception is of limited importance, since the radionuclides can be subsequently transferred to the soil by leaching and wash-off and by decay when the vegetation dies back or sheds its leaves. For, ^7Be , however, this interception can be more important,

since the short half-life of the radionuclide can mean that a significant proportion of the fallout input may be intercepted and subsequently lost from the vegetation canopy through radioactive decay and therefore never reach the soil surface. Work in forested areas in Maine, USA, has indicated that as much as 50% of the fallout input could be held in the forest canopy [85] and a study of ^7Be interception by grass undertaken at Bologna, Italy similarly indicated that more than 50% of the total inventory was held in the grass layer above the soil [86]. Kaste et al. [66] also suggest that soils with a grass cover are frequently characterized by a higher total areal activity density than bare soils, pointing to the role of the grass in trapping atmospheric aerosols. This interception has important implications for the use of ^7Be as a tracer of soil redistribution, since in areas with a vegetation canopy, it is likely to result in considerable local-scale variability in the soil inventory and it may prove impossible to establish a reference inventory for the soil itself which can be used to identify areas of erosion (i.e., reduced ^7Be inventories) and deposition (i.e., areas with increased ^7Be inventories).

Because of the rapid and strong adsorption of the fallout input by the surface horizons of the soil, the total inventory or areal activity density (Bq m^{-2}) measured at a reference site, defined as a stable undisturbed site which has experienced neither erosion nor deposition, will directly reflect the prior fallout input to the site, assuming that the transfer to the soil has not been influenced by a vegetation canopy as discussed above. Figure 5 provides a schematic representation of the variation of that inventory over the timescale relevant to the use of a particular fallout radionuclide as a tracer. For ^{137}Cs and ^{210}Pb , with half-lives extending to decades, this timescale covers ca. 50–100 years, whereas for ^7Be , with a much shorter half-life, it extends over ca. 6 months. For ^{137}Cs (Fig. 5a), the key feature of the variation of the inventory over the period of interest is the absence of ^{137}Cs in the soil prior to the early 1950s, the subsequent rapid increase in the inventory until the mid-1960s, due to the bomb tests during the period extending from the mid-1950s until 1963, and the subsequent decline in the inventory since the mid-1960s, when loss from the inventory due to radioactive decay exceeded the residual fallout input from the bomb tests. In the absence of new ^{137}Cs fallout, the inventory will continue to progressively decline to



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 5

A schematic representation of the interaction between the fallout input and the (a) ^{137}Cs , (b) $^{210}\text{Pb}_{\text{ex}}$ and (c) ^7Be inventory in the soil of a reference site over the timescale relevant to the use of the individual fallout radionuclides as tracers of soil and sediment redistribution

the point where it will fall below the level of detection and the radionuclide can no longer be used as a tracer. This point is likely to be reached earlier in equatorial regions and in the southern hemisphere, where inventories are lower. In locations where an additional input of ^{137}Cs occurred in 1986, as a result of the Chernobyl accident, this input would have caused an increase in the inventory and a temporary halt to the decline. However, since Chernobyl fallout was short-lived, the declining trend was soon restored.

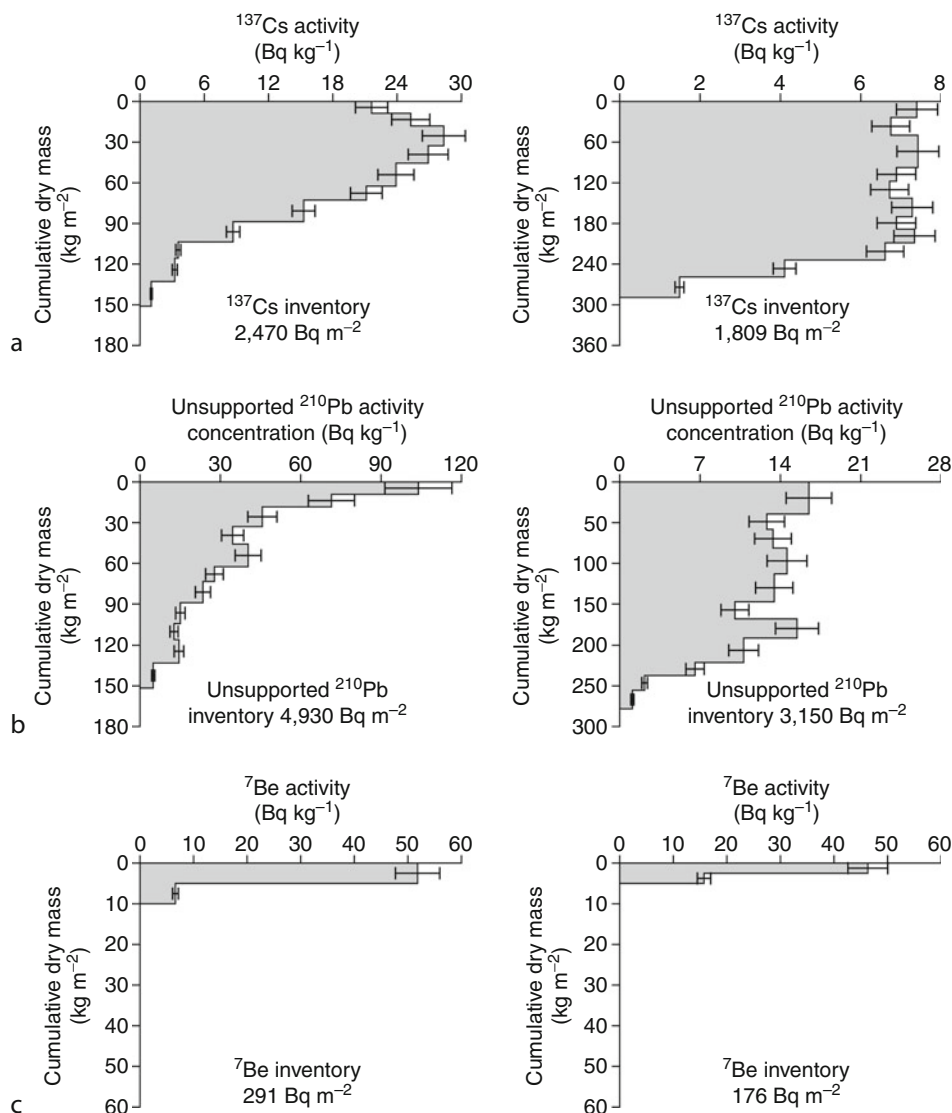
For $^{210}\text{Pb}_{\text{ex}}$, the pattern of variation of the inventory is very different from that of ^{137}Cs and the inventory can be seen as essentially constant through time (Fig. 5b), with new fallout inputs balanced by losses caused by radioactive decay. For a reference site with

a mean annual fallout input of $250 \text{ Bq m}^{-2} \text{ year}^{-1}$, the steady-state inventory will be $\sim 8,169 \text{ Bq m}^{-2}$. In reality, there will be small variations in the total inventory due to inter-annual variations in the fallout input, but since the annual fallout input represents only a relatively small proportion of the total inventory (i.e., ca. 3%), the effect of these fluctuations will be small. In the case of ^7Be , however, the magnitude and timing of the fallout inputs will exert a key influence on the variation of the inventory through time, due to the short half-life of this radionuclide. In the absence of further fallout, the inventory would decline to near-zero within about 6 months and the temporal pattern of fallout inputs will dominate the variation of the inventory through time. A timescale of ca. 6 months

has therefore been used in Fig. 5c to represent the key features of the variation of the ^7Be inventory through time.

When considering in more detail the storage or the longer-term fate of fallout radionuclides within the soil, there is a need to distinguish undisturbed and cultivated soils. In the former, redistribution of the fallout input within the soil profile will be governed

primarily by physicochemical diffusion and migration processes, including biological activity, whereas in the latter case, tillage will commonly mix the upper horizons of the soil and cause the radionuclide activity to be relatively uniformly distributed through the plow layer. Figure 6 presents typical ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$ and ^7Be depth distributions for undisturbed (i.e., permanent pasture) and adjacent cultivated area documented by the author



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 6

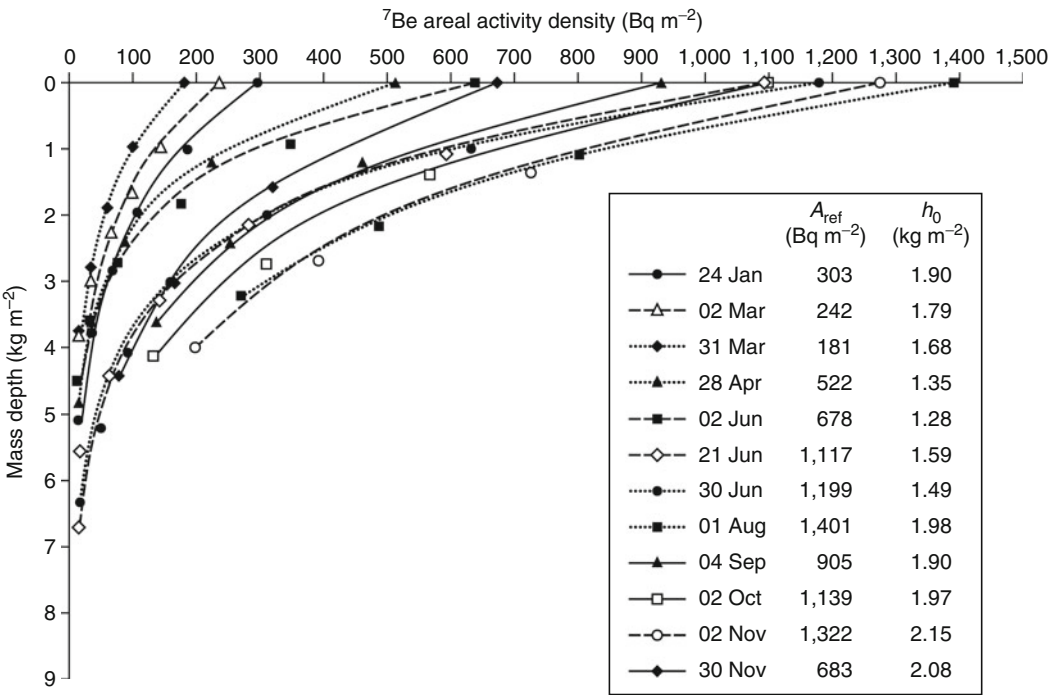
Characteristic depth distributions of ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$, and ^7Be in undisturbed pasture soils (*left*) and cultivated soils (*right*) in Devon, UK. The precision of the radionuclide measurements at the 95% level of confidence is indicated by the error bars

and his coworkers for a site near Crediton in Devon, UK. Depth distributions of similar shape are likely to be found in different areas of the world, although the total inventory will vary. In Fig. 6, the depth scale has been expressed as a mass depth (kg m^{-2}), since this removes the need to take account of variations in bulk density within the upper levels of the soil in influencing the precise form of the depth distribution. For ^{137}Cs , the depth distribution for an undisturbed soil shown in Fig. 6a is characterized by the typical exponential decline in activity with depth that has been widely reported for undisturbed soils. In this case, ca. 80% of the total inventory is found within the upper 10 cm of the soil profile. This depth distribution can be readily accounted for by the surface input of ^{137}Cs as fallout and its subsequent downward movement by diffusion-like processes. These processes include biological activity (e.g., earthworms), which cause mixing. In many undisturbed soil profiles, such as that illustrated here, the peak activity is found just below the surface, rather than at the surface, and this can be seen as reflecting the absence of significant fallout since the 1980s (in the absence of Chernobyl fallout), biological activity moving soil from depth to the surface, and the accumulation of dead organic material at the surface. The depth distribution of $^{210}\text{Pb}_{\text{ex}}$ for an undisturbed soil, illustrated in Fig. 6b, is, as might be expected, very similar to that of ^{137}Cs . However, in this case, the fallout input is continuous and the maximum activity is commonly found at the surface, where it is maintained by the fresh fallout. The depth distribution of ^7Be associated with an undisturbed soil shown in Fig. 6c differs substantially from that of the other two fallout radionuclides, in that all the inventory is contained within the upper ca. 10 mm of the soil and most is within the upper 5 mm. This situation reflects the continuing input of ^7Be fallout to the surface and the short half-life of the radionuclide, which means that any ^7Be slowly moving deeper into the soil disappears quite rapidly due to decay.

In the case of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$, the depth distributions illustrated for an undisturbed soil in Fig. 6 can be viewed as essentially constant over a timescale of several years. The $^{210}\text{Pb}_{\text{ex}}$ profile is likely to remain constant over a longer timescale, as the fallout input and downward movement are balanced by decay. In the longer

term, the ^{137}Cs depth distribution would have changed, through time, reflecting the beginning of the fallout input in the 1950s, when most of the radiocesium would have been found near the surface, the subsequent progressive downward movement of the radionuclide, the cessation of significant fallout in the 1980s, and the gradual reduction in the total inventory after the mid-1960s, when loss by decay started to exceed the input of new fallout. Existing experimental evidence, based on the artificial application of radiocesium to bare soil surfaces, indicates that fresh fallout is initially distributed within the top few millimeters of the soil. This situation is essentially replicated by the ^7Be depth distribution, since decay removes the evidence of medium-term downward movement. Because of its short half-life, the ^7Be depth distribution will be more dynamic, reflecting short-term changes in the total inventory in response to fresh fallout and its subsequent redistribution within the shallow upper soil layer, as well as ongoing decay. Figure 7 illustrates the evolution of the depth distribution of ^7Be within a bare undisturbed soil in an area of cleared forest near Valdivia in southern Chile, from summer through the wet season (autumn and winter) to the following spring [87]. Over the period of measurement, the total inventory varies over nearly an order of magnitude from 181 to $1,401 \text{ Bq m}^{-2}$. In Fig. 7, the depth distributions are represented by the total inventory recorded above a given depth and the individual distributions show a clear exponential form. The h_0 values listed represent the relaxation depths of the exponential distributions and increased values reflect greater penetration. During the initial period, which represents the end of the dry season, the inventory declines, but it subsequently increases at the onset of the wet season. At this stage, the new input of fallout to the surface causes an increase in the inventory held close to the surface and the h_0 values decline. As the wet season continues, however, the depth distribution deepens and the h_0 values increase to reflect the greater proportion of ^7Be held at depth.

As shown in Fig. 6, the depth distributions of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ in cultivated soils are dominated by the influence of soil mixing caused by tillage. The profiles are characterized by a fairly uniform radionuclide activity in the plow layer, which is well mixed by tillage, but the activity rapidly declines below the plow depth.



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 7
Depth distributions of ⁷Be recorded in a bare forest soil at a reference site near Valdivia in southern Chile over the seasonal cycle. The inventories (A_{ref}) and relaxation depths (h_0) associated with the individual depth distributions are listed in the key (Based on [87])

Most profiles evidence a small “tail” extending below the plow depth, reflecting some limited downward movement of the radionuclide to greater depth. The ¹³⁷Cs and ²¹⁰Pb_{ex} profiles for cultivated soils portrayed in Fig. 6 were taken from eroding areas within a cultivated field, and the total inventories are less than those associated with the equivalent cores from an adjacent area of undisturbed grassland. The reduced inventories reflect the progressive loss of soil containing ¹³⁷Cs or ²¹⁰Pb_{ex} from the surface of the profile by erosion. As erosion proceeds, both the inventory and the ¹³⁷Cs activity in the plow layer will decline, as plowing incorporates subsoil containing little ¹³⁷Cs or ²¹⁰Pb_{ex} into the plow layer. In the absence of erosion, the depth distribution would be characterized by a total inventory similar to that associated with the undisturbed stable site. The shape of the profile would also be similar to those presented in Fig. 6 for cultivated soils, but the activity within the plow layer would be greater.

In contrast to the ¹³⁷Cs and ²¹⁰Pb_{ex} profiles depicted in Fig. 6, the depth distribution of ⁷Be for

the cultivated field is almost identical to that recorded for the undisturbed area. The reason is that the ⁷Be fallout accumulates in the upper few millimeters of the soil and when the field is plowed, the action of mixing the accumulated ⁷Be into the plow layer reduces the activity to below the level of detection. After plowing and cultivation, new fallout will again accumulate near the surface and as time progresses and the inventory grows, the depth distribution will closely resemble the depth distribution associated with the undisturbed area. The reduced ⁷Be inventory associated with the depth distribution for the cultivated field relative to the adjacent undisturbed field could reflect the timing of the prior plowing and the limited time available for the inventory to be restored. However, in this instance, the plowing occurred many months previously and the reduced inventory reflects the loss of surface soil by erosion. This situation is confirmed by the reduced depth to which ⁷Be is found.

Where soil redistribution within the landscape caused by erosion produces areas of deposition, the

depth distributions of the fallout radionuclides associated with such depositional areas will differ from those for stable and eroding areas presented in Fig. 6. Assuming that the mobilized sediment which is subsequently deposited contains ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$, or ^7Be , by virtue of its source within the landscape, the deposition of sediment, for example at the foot of a slope or on a river floodplain, will cause the inventory to increase and the depth distribution to change due to upward extension and thus an increase in the depth of soil or sediment containing the radionuclide. Figure 8 provides some examples of such changes associated with a river floodplain site in Devon, UK. In this case, the depth distributions were documented for both a point on the river floodplain and an adjacent area of undisturbed pasture above the level of inundation, with no evidence of either erosion or deposition. In both cases, the soils were uncultivated and the profiles are therefore undisturbed by tillage mixing. In the case of the ^{137}Cs depth distributions, the profile from the floodplain has an inventory that is considerably greater than the profile from the area above the level of inundation, as a result of deposition of sediment containing ^{137}Cs . The profile shape also provides clear evidence of deposition. The shape of the ^{137}Cs depth distribution shown for the floodplain reflects the progressive reduction of the ^{137}Cs content of the sediment sources within the catchment, and thus also of the deposited sediment, as a result of the ongoing loss of ^{137}Cs through erosion.

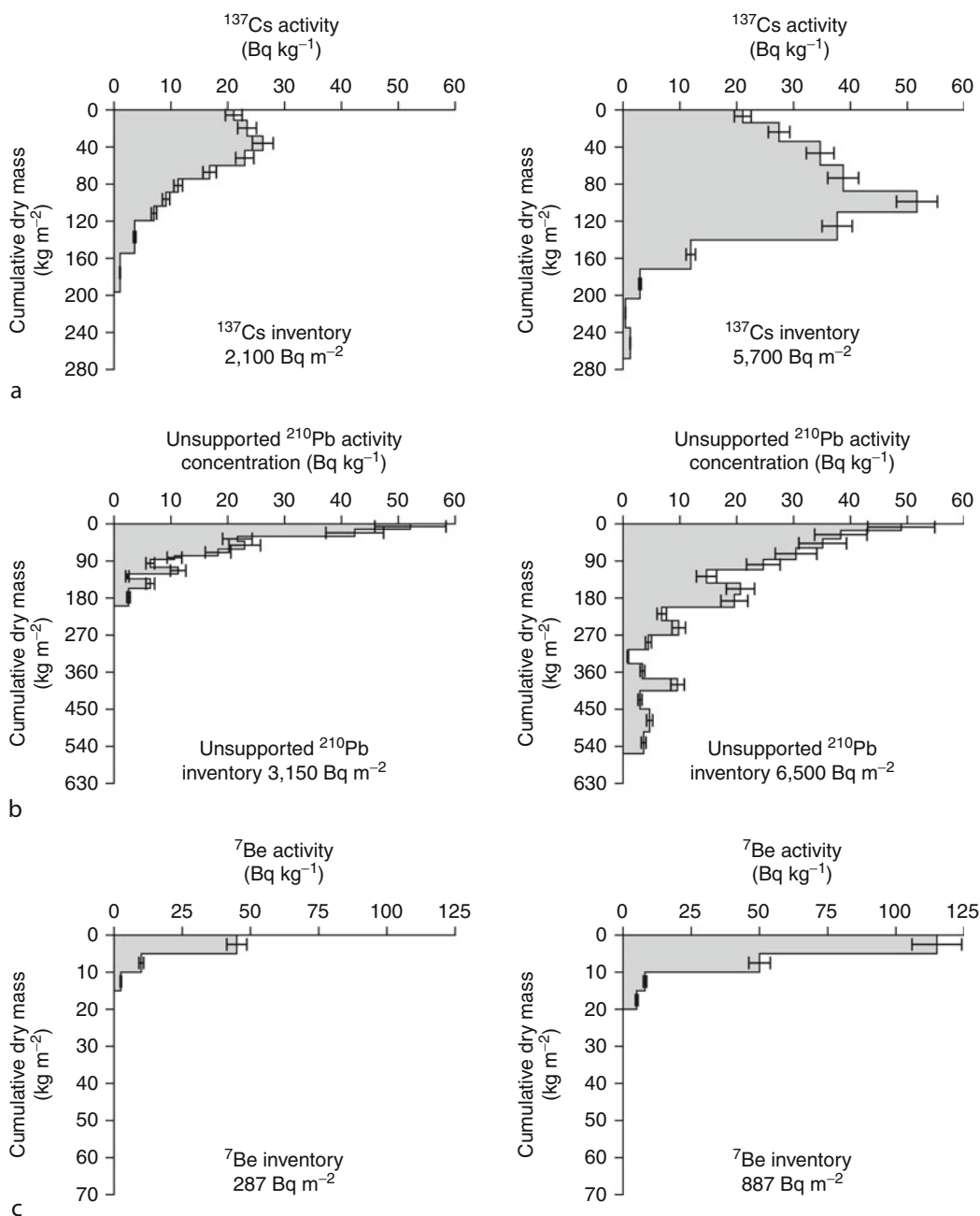
As with ^{137}Cs , the total inventory associated with the $^{210}\text{Pb}_{\text{ex}}$ depth distribution from the depositional site on the floodplain shown on Fig. 8 is substantially greater than that of the $^{210}\text{Pb}_{\text{ex}}$ profile for the undisturbed site above the level of inundation. This again reflects the input of $^{210}\text{Pb}_{\text{ex}}$ to the floodplain profile from both fallout to the floodplain surface and the deposition of sediment containing $^{210}\text{Pb}_{\text{ex}}$. The depositional environment is again further confirmed by the substantially greater depth to which $^{210}\text{Pb}_{\text{ex}}$ is found in the floodplain profile. Since the fallout input of $^{210}\text{Pb}_{\text{ex}}$ to both the upstream catchment and the floodplain surface is essentially continuous, there is no reduction in activity toward the surface of the floodplain profile. The rate of decrease of the $^{210}\text{Pb}_{\text{ex}}$ activity with depth will reflect the rate of sediment accretion and the rate of decay of the $^{210}\text{Pb}_{\text{ex}}$ contained in the buried sediment. Since the latter is known, it

is possible to estimate the rate of sediment accretion from the rate of decrease of activity with depth in the profile (see section “[Estimating Sedimentation Rates and Establishing Chronologies in Depositional Sinks](#)”). Comparison of the two ^7Be depth distributions presented in Fig. 8 again provide clear evidence of deposition, as reflected by the increased inventory and depth of the ^7Be profile from the accreting floodplain, when compared with the profile from the site above the level of inundation. Both these features of the floodplain profile indicate that the deposition occurred only a short time prior to the time that the ^7Be depth distributions were documented, since in the absence of new sediment deposition, any difference between the two ^7Be profiles would soon disappear. The higher ^7Be surface activity associated with the floodplain profile reflects the two sources of ^7Be input to the surface, namely, direct fallout and deposition of sediment containing ^7Be mobilized from the upstream catchment.

The fallout radionuclide profiles from a depositional site illustrated in Fig. 8 relate to an area of undisturbed permanent pasture on a river floodplain. Similar profiles will be found in other depositional sinks, such as areas of deposition at the foot of a slope, where the soil surface will receive both direct fallout and an input of radionuclide-bearing sediment eroded from upslope. In many instances, depositional sites will also be cultivated. In cultivated areas, the depth distribution of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ associated with depositional areas will again be characterized by increased inventories and the occurrence of the two radionuclides to greater depths. However, the action of tillage will commonly introduce greater homogeneity into the depth profile and in many cases, the magnitude of the increase in inventory rather than the shape of the depth distribution will provide the key evidence as to the rate of deposition.

Using Fallout Radionuclides as Tracers to Investigate Soil and Sediment Redistribution within the Landscape

A sound understanding of the input and redistribution of ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$, and ^7Be within the landscape provides the key to using these three radionuclides to estimate rates of soil and sediment redistribution.



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 8

Characteristic depth distributions of ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$ and ^7Be from an undisturbed pasture site (*left*) and from a neighboring regularly inundated river floodplain (*right*) in Devon, UK. The precision of the radionuclide measurements at the 95% level of confidence is indicated by the error bars

Their distinctive behavior at erosional and depositional sites, in particular, provides a basis for quantifying erosion and deposition rates and discriminating fine sediment sources. Further discussion of the principles

involved will consider: firstly, quantification of soil redistribution within the landscape by soil erosion, secondly, the estimation of sedimentation rates in depositional sinks such as lakes, reservoirs, and river

floodplains and establishing chronologies for the associated sediment deposits, and thirdly, discriminating the potential sources of fine sediment in catchments and river basins. In each case, there is a need to take account of differences between the three fallout radionuclides in terms of the temporal pattern of fallout and the timescale of interest and the consequent need for different approaches.

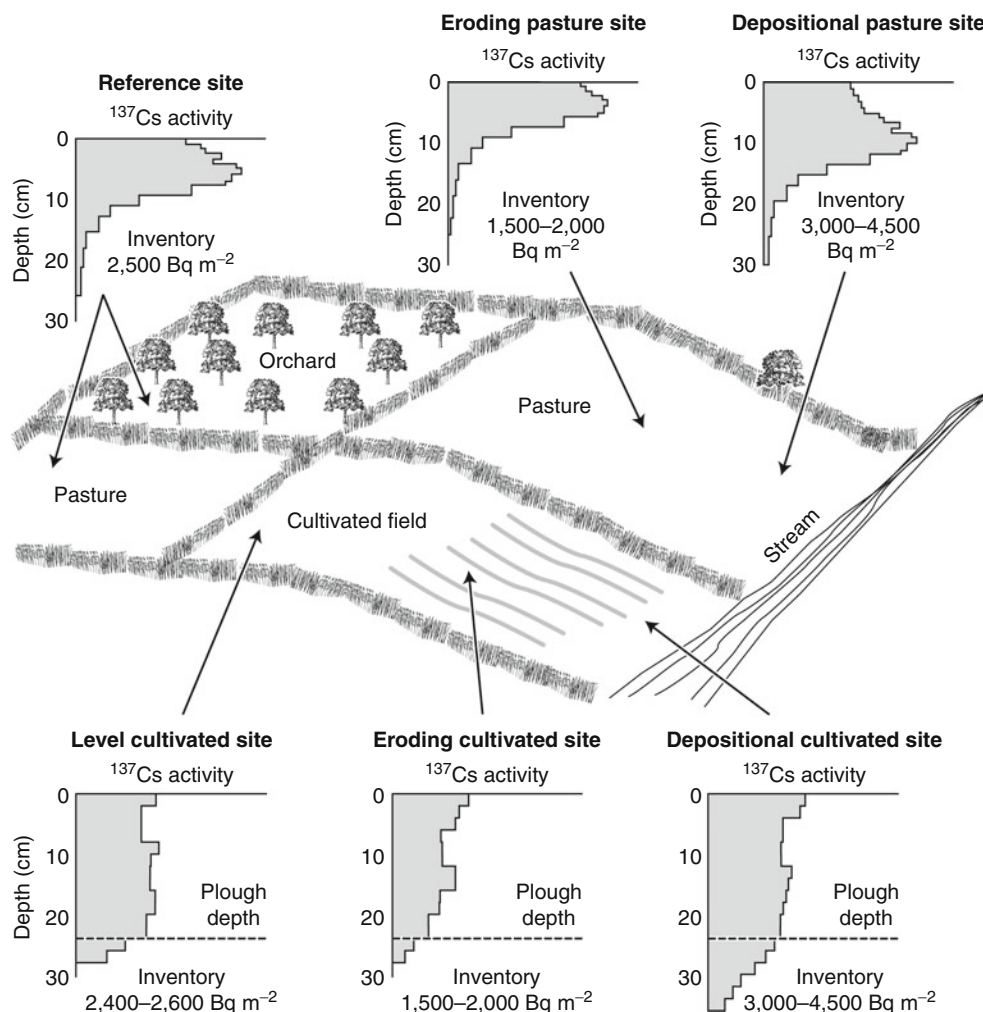
Quantification of Soil Redistribution Within the Landscape by Soil Erosion

As demonstrated by Figs. 6 and 8, both the total inventory and the depth distribution of ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$, and ^7Be in soil or sediment can provide information on past soil redistribution at that point. Because the inventory can generally be obtained with a single measurement, whereas a substantial number of measurements will be needed to establish the depth distribution, and the number of samples that can be measured is frequently limited by detector availability and the count times involved (see section “Laboratory Procedures and Techniques”), attention usually focuses on inventory measurements. In most studies, the approach taken involves establishing a reference inventory, which represents the inventory associated with an undisturbed location experiencing neither erosion nor deposition, and comparing the inventories obtained for other sampling points with this value. Where they are less than the reference inventory, it is assumed that erosion has occurred and, where they exceed the reference inventory, this is assumed to reflect deposition. In the case of ^{137}Cs , the soil redistribution resulting in the reduced or increased inventory is assumed to have occurred over the period since the main period of ^{137}Cs fallout in the late 1950s and early 1960s, whereas for $^{210}\text{Pb}_{\text{ex}}$ and ^7Be , the period is defined by considering the half-life of the radionuclide and the extent to which the current inventory will reflect past fallout and redistribution of that fallout. It is important that this comparison should take account of the uncertainty associated with both the precision of the laboratory measurements of the radionuclide activity used to establish the inventories of the sampling points and any uncertainty associated with the value used for the reference inventory. The latter will reflect both the measurement precision and the uncertainty introduced by small-scale spatial

variability of the soil inventory and the sampling procedure (see section “Field Techniques for Sample Collection”).

In a few studies involving ^{137}Cs , a different repeat sampling approach has been employed and soil redistribution rates are estimated by comparing the inventories measured on two occasions at the individual sampling points [88]. The principle involved is very similar to the use of a reference inventory, in that a reduction in inventory over time is assumed to reflect erosion, whereas an increase is taken to indicate deposition. Although this alternative approach has the advantage that the period during which soil redistribution processes have been operating to cause the changing inventory is precisely defined and the need to establish the local reference inventory is removed, it also has several important limitations which are likely to restrict its application. The first is the need for repeat measurements and thus the inability to obtain results quickly and without preplanning. Studies where the approach has been used have generally been able to make use of data fortuitously provided from earlier sampling programs, but these will rarely be available. Secondly, there is a need to ensure that the period elapsed between the two surveys is of sufficient duration to produce a significant reduction or increase in the inventory, bearing in mind the uncertainty associated with the measured inventories, which will reflect both analytical precision and variability introduced by sampling. Finally, relocating sampling points may prove difficult. Even in areas where erosion and deposition rates are relatively high, it has been estimated that a period of 10 years between sampling campaigns may be required to ensure that there is a significant difference between the two sets of measurements. Where soil redistribution rates are relatively low, the length of the period will increase. This contribution will focus on the standard approach using a reference inventory.

Use of ^{137}Cs Measurements Figure 9 provides a schematic representation of the principles involved in using ^{137}Cs measurements to estimate soil redistribution rates. Here, an orchard is used as a reference site and the cores used to define the reference inventory and the characteristic depth distribution are collected between the trees where the presence of the trees will have had



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 9

A schematic representation of the principles involved in the use of ^{137}Cs measurements to estimate soil redistribution rates

little or no influence on fallout receipt. Eroding areas within both pasture and cultivated areas will be characterized by reduced inventories. Similarly, areas of deposition are reflected by increased inventories. The depth distributions provide further evidence of the incidence and magnitude of erosion and deposition. As indicated above, the soil redistribution reflected by the departure of the measured inventory values from the reference value is assumed to have occurred over the period between the main period of ^{137}Cs fallout in the late 1950s and 1960s and the present. Although information on ^{137}Cs inventories and depth

distributions, such as that shown in Fig. 9, provides *qualitative* information on the magnitude and spatial distribution of soil redistribution rates within the landscape, for most purposes, there will be a need for quantitative estimates of soil redistribution rates. To obtain such estimates, it is necessary to apply what have become known as “conversion models” to the measured inventory values.

Reviews of the various conversion models used with ^{137}Cs measurements (e.g., [89–91]) commonly distinguish between *empirical* models and *theoretical* models. Empirical models, as might be expected, are based on

the use of empirical measurements to establish the relationship between the magnitude of the loss or gain of ^{137}Cs inventory relative to the reference inventory and the soil redistribution rate. Such data are difficult to obtain, but there have been some attempts to use data from a range of erosion plots for this purpose, as in Australia, where relationships between rate of soil erosion and the reduction in ^{137}Cs inventory have been established for both cultivated and uncultivated soils [92]. In this case, the measured soil loss from the plots Y ($\text{t ha}^{-1} \text{ year}^{-1}$) was related to the reduction in ^{137}Cs inventory within the plot, X (%), using a function of the form

$$Y = aX^b \quad (1)$$

However, the development and use of such empirical models involves several important problems. Perhaps most importantly, any relationship developed will be time specific. This constraint is particularly important. For example, loss of 20% of the inventory by 1980 would reflect a much higher erosion rate than loss of 20% of the inventory by 2000 and a relationship established using data for 1980 would overestimate erosion rates if applied to recent measurements of the reduction in inventory. Equally, it is difficult to use this approach to estimate deposition rates, although, in some investigations, it has been assumed that the equation can simply be rearranged using the same constant and coefficient to estimate deposition.

Because of the lack of data that can be used to establish empirical models and the inherent limitations noted above, most of the conversion models that have been successfully used for estimating soil redistribution rates from ^{137}Cs measurements fall into the second group, namely, theoretical models. By definition, such models use existing understanding of the behavior and fate of fallout inputs in response to soil redistribution processes to derive a theoretical relationship between soil redistribution rate and the loss or gain of inventory relative to the reference inventory. The proportional model [21, 93] provides a very simple example of such a model that has been quite widely used for cultivated soils. In this case, the plow depth H (m) is determined and the depth of soil lost, as a proportion of H , is estimated by simply assuming that this proportion will be the same as X (%), the proportion of the inventory that has been lost. If the bulk density

of the soil B (kg m^{-3}) and the period over which erosion T (years) has occurred are known, the depth of soil lost can be used to estimate the erosion rate Y that is,

$$Y = 10 \frac{BdX}{100T} \quad (2)$$

Although attractive in its simplicity, the basic assumptions underlying the proportional model are flawed, since it takes no account of the fact that as soil is eroded from the surface of the plow layer in a cultivated field, the depth of the plow layer will be maintained by incorporation of subsoil from below the original plow layer. Therefore, if a depth equivalent to the depth of the plow layer was removed by erosion, the soil would still contain a significant proportion of the original inventory and not an inventory of zero. Use of the proportional model will therefore *underestimate* erosion rates, and it has been shown [91] that with high rates of soil loss (e.g., loss of ca. 60% of the original inventory), the estimated erosion rate is likely to underestimate the erosion rate by ca. 40%. The simple assumptions underlying the proportional model relate to an eroding soil and the model does not explicitly provide a means of estimating deposition rates, where the measured inventory exceeds the reference inventory. However, in several studies, the equation has been extended to estimate the deposition rate, by again assuming that the depth of deposition represents a proportion of H , equivalent to the percentage increase in inventory, relative to the reference inventory. Use of this procedure is, however, likely to underestimate the deposition rate, since the deposited soil eroded from upslope and deposited at a sampling point is likely to be characterized by a lower ^{137}Cs activity than that assumed by the model.

Other attempts to develop conversion models have aimed to provide a more realistic representation of the behavior of ^{137}Cs in the soil at locations characterized by erosion and deposition [90, 91]. In providing a brief overview of some of these models, it is useful to distinguish between those developed for cultivated soils and those developed for uncultivated soils, for example, soils found on pasture or rangeland. In most circumstances, reduction of the soil inventory by a given percentage will reflect a much higher erosion rate for a cultivated soil, than for an uncultivated soil, since

the incorporation of the ^{137}Cs inventory into the plow layer means that a much greater depth of soil will need to be eroded to reduce the soil inventory by a given amount. For an uncultivated soil, most of the ^{137}Cs inventory is likely to be found close to the surface and reduction of the inventory by a given proportion is likely to reflect a much lower erosion rate than for a cultivated soil.

Since they have attracted most attention in terms of conversion model development, attention will be directed first to the use of theoretical conversion models developed for use with cultivated soils. Most of these models, which aim to overcome the fundamental inadequacy of the proportional model, use a mass balance approach and attempt to model the changes in the ^{137}Cs content of the plow layer in response to fallout inputs, radioactive decay, loss of ^{137}Cs from the surface by erosion, and incorporation of soil containing no ^{137}Cs from beneath the plow depth by ongoing plowing, over the period since the commencement of ^{137}Cs fallout. By running the model with a range of erosion rates, it is possible to establish a relationship between erosion rate and the degree of reduction of the ^{137}Cs inventory relative to the reference inventory. The basic form of the mass balance model for an eroding site can be represented as follows:

$$\frac{dA(t)}{dt} = I(t) - \left(\lambda + \frac{R}{d_m} \right) A(t) \quad (3)$$

where $A(t)$ = cumulative inventory (Bq m^{-2}), t = time since onset of ^{137}Cs fallout (years), R = erosion rate ($\text{kg m}^{-2} \text{ year}^{-1}$), d_m = average plow depth expressed as a mass depth (kg m^{-2}), λ = decay constant for ^{137}Cs (year^{-1}), and $I(t)$ = annual deposition flux at time t ($\text{Bq m}^{-2} \text{ year}^{-1}$).

For many locations, detailed information on the annual ^{137}Cs deposition is unlikely to be available and in this situation, the fallout record from a representative station in the same hemisphere (see Fig. 2) can be scaled to match the measured reference inventory for the study site. In order to estimate deposition rates R' ($\text{kg m}^{-2} \text{ year}^{-1}$) using the same mass balance approach, there is a need to take account of the magnitude of the increase in the ^{137}Cs inventory. In this case, the deposition rate will reflect the ^{137}Cs

activity in the deposited sediment, $C_d(t')$ (Bq kg^{-1}) according to:

$$R' = \frac{A(t) - A_{\text{ref}}}{\int_{t_0}^t C_d(t') e^{-\lambda(t-t')} dt'} \quad (4)$$

where $C_d(t')$ = the ^{137}Cs activity of deposited sediment in year t' (Bq kg^{-1}).

The ^{137}Cs activity of deposited sediment can be estimated as the weighted mean ^{137}Cs activity of sediment mobilized from the upslope contributing area (i.e., from the data provided by eroding points upslope).

The general framework provided by Eqs. 3 and 4 has been extended by various workers to take account of other processes and factors that can be expected to influence the relationship between the erosion rate and the percentage reduction in the inventory relative to the reference inventory (e.g., [90, 94, 95]). These include the fate of fresh fallout which can be removed by erosion prior to incorporation into the plow layer by tillage, the particle size selectivity of erosion and deposition processes and the preferential association of ^{137}Cs with the finer fractions of the soil and with organic matter, and the role of tillage erosion in influencing ^{137}Cs redistribution. In the latter case, it has proved possible to develop conversion models that permit separation of water erosion and tillage erosion and provide separate estimates of each. In an overview of the conversion models available for cultivated soils, Walling and He [90, 91] provide further details of a simple mass balance model which assumes that all fallout was received in 1963, a more comprehensive mass balance model incorporating the fate of fresh fallout and the size selective nature of erosion and deposition and the further extension of that model to incorporate the effects of ^{137}Cs redistribution by tillage.

In the case of uncultivated soils, the conversion model needs to take account of the shape of the depth distribution, since this will influence the relationship between erosion rate and the depletion of the total ^{137}Cs inventory. Two main approaches have been used in developing models for uncultivated soils [91]. The first, which is often termed the profile distribution model, is based on the assumption that the ^{137}Cs depth distribution in most uncultivated soils is

exponential in form and can be characterized by a relaxation depth h_0 (kg m^{-2}). The greater the value of h_0 the greater the depth to which the ^{137}Cs extends into the soil. If the value of h_0 is determined for the reference site and it is assumed that at eroding sites, loss of surface soil will cause the exponential depth distribution to be truncated, comparison of the inventory at the sampling point with the reference inventory provides a means of estimating the depth of soil which must have been removed to reduce the inventory by the documented amount. The erosion rate can be estimated by dividing the total depth of soil removed by the number of years elapsed since 1963, the year of peak fallout, namely,

$$Y - \frac{10}{t - 1963} \ln \left[\left(1 - \frac{X}{100} \right) \right] h_0 \quad (5)$$

where t = the year of sample collection.

Where the measured inventory at a sampling point exceeds the reference inventory, indicating that deposition has occurred, the deposition rates can be estimated from the magnitude of the increase in inventory and the ^{137}Cs concentration of deposited sediment, using the same approach as used for cultivated soils in Eq. 4.

Use of the profile distribution model assumes that the exponential depth distribution found at the reference site would have developed at the eroding point by 1963 and was subsequently truncated by erosion. This is clearly an oversimplification, since the depth distribution will evolve through time as a result of the time-variant fallout input, the progressive downward movement of the ^{137}Cs fallout received at the soil surface, and the removal of some fresh fallout from the soil surface by erosion shortly after its deposition. As a result, estimates of rates of soil loss provided by the profile distribution model are likely to overestimate erosion and deposition rates. However, a study aimed at testing the model within a small catchment in southern Italy [96], by comparing the measured sediment yield from the catchment with the estimate of mean annual net soil loss provided by the application of the profile distribution model to ^{137}Cs measurements obtained from the catchment, demonstrated good agreement between the two independent results.

Uncertainties regarding the assumptions implicit in the profile distribution model have encouraged the development of an alternative approach, commonly referred to as the diffusion and migration model [91]. This model involves a mass balance approach taking account of the time-variant fallout input and the slow downward movement of the ^{137}Cs fallout received at the surface. As such, it takes account of the evolution of the ^{137}Cs depth distribution through time, rather than assuming a constant form. This downward movement is governed by two lumped parameters loosely referred to as the diffusion and migration coefficients, which can be estimated from the form of the depth distribution found at the reference site. With this model, it is again possible to estimate the deposition rate associated with a sampling point where the inventory exceeds the local reference inventory. The approach taken is similar to that represented in Eq. 4. Since this model more closely replicates existing understanding of the post-fallout redistribution of ^{137}Cs in undisturbed soils and its interaction with erosion, it is likely to provide more reliable estimates of soil redistribution rates in areas with uncultivated soils.

Although most studies employing ^{137}Cs to investigate erosion and soil redistribution rates have focused on water erosion, there have been some studies that have attempted to extend its application to wind erosion. These studies have employed broadly the same approach of comparing the reference inventory at an undisturbed site unaffected by erosion or deposition with the inventories measured at a range of sampling points. Points with reduced inventories are assumed to represent eroding points, and points with increased inventories are assumed to represent points where deposition has occurred. In most instances, conversion models based on those developed for water erosion have been used, since the mechanisms involved are similar, that is, removal of soil from the surface and addition of sediment to the original surface by deposition. As with water erosion, a range of different approaches have been used in the models. These include the proportional model [97, 98], mass balance models [99, 100], and a profile distribution model [100, 101]. However, the available evidence suggests that wind erosion introduces greater complexity into the post-fallout redistribution of ^{137}Cs and the use of

^{137}Cs measurements to estimate soil and sediment redistribution rates. Important issues include the potential importance of selective winnowing of fines by deflation processes and the possibility that reference sites could be influenced by dust accumulation, which could increase the measured inventory [102].

As indicated above, estimates of soil redistribution rates obtained using ^{137}Cs measurements generally relate to the period extending from either the onset or the peak period of fallout in the 1950s or 1960s to the time of sampling. In many respects, the lengthy period covered by the measurements represents an advantage of the ^{137}Cs approach, because it integrates the natural temporal variability of soil erosion rates. However, it can also represent a limitation of the approach, in that it generates an average erosion rate for the period and it is not possible to document changes in erosion rates resulting from land use change. However, where land clearance for agriculture and the onset of significant erosion postdates the main period of ^{137}Cs fallout, the conversion models can be modified to provide an estimate of soil redistribution relating only to the post-clearance period. In another specialist application, it proved possible to produce estimates of erosion rates associated with two contrasting land management practices that characterized the period of interest [103, 104]. In this case, it was possible to reconstruct the erosion rates occurring during a period of conventional cultivation extending from 1954 to 1986 and a subsequent period under no-till management extending from 1986 to 2003. By measuring the current ^{137}Cs activity of the former plow layer associated with the period of conventional tillage, it was possible to reconstruct the inventory of a sampling point at the end of this period and thus estimate the erosion rate under conventional tillage using a mass balance conversion model. Similarly, the current depth of the original plow layer was estimated from measurements of the current total ^{137}Cs inventory and the current ^{137}Cs activity of the plow layer at individual sampling points and this value was compared with the plow depth during the period of conventional cultivation to provide an estimate of the depth of soil eroded during the period under no-till management. To use this approach, it is important that the overall period covered by the ^{137}Cs measurements (i.e., ca. 50 years) should include substantial periods under both

management practices and thus that the change in management should coincide approximately with the middle of the period covered by the ^{137}Cs measurements.

Use of $^{210}\text{Pb}_{\text{ex}}$ Measurements The use of $^{210}\text{Pb}_{\text{ex}}$ measurements to investigate rates and patterns of soil redistribution within the landscape involves similar principles to those involved in the use of ^{137}Cs and shown schematically in Fig. 9. As with ^{137}Cs , the inventories and depth distributions documented at individual measuring points can be compared with equivalent information obtained from an undisturbed reference site and a conversion model is used to estimate soil redistribution rates. The key differences between the use of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ measurements relate to the contrasting nature of the fallout input and the shorter half-life of $^{210}\text{Pb}_{\text{ex}}$ (22.3 years vs 30.17 years). The essentially continuous nature of $^{210}\text{Pb}_{\text{ex}}$ fallout means that, unlike the situation with ^{137}Cs , there is no well-defined period (e.g., from the commencement of fallout to the time of sampling) for which soil redistribution rates are documented. The $^{210}\text{Pb}_{\text{ex}}$ inventory measured at a sampling point will reflect the interaction between the fallout input and soil redistribution processes over an extended period. This period will reflect the half-life of the radionuclide, such that the impact of past soil redistribution on the present day inventory will become progressively less as the period extends back into the past. It is frequently indicated that $^{210}\text{Pb}_{\text{ex}}$ measurements can be used to estimate the erosion occurring over a period representing about five half-lives and thus over the past ca. 100 years. In principle, this is correct, but it must be recognized that erosion occurring 100 or even 50 years ago will have much less influence on the current inventory than erosion occurring in the past 10 or 20 years. Thus, although $^{210}\text{Pb}_{\text{ex}}$ measurements will reflect erosion and soil redistribution taking place over a longer period than ^{137}Cs , the measurements are likely to be more sensitive to recent erosional activity than ^{137}Cs measurements.

Walling and his coworkers [38, 105] describe several conversion models which have been developed for deriving estimates of soil redistribution rates, based on the degree of reduction or increase in the $^{210}\text{Pb}_{\text{ex}}$ inventory relative to the reference value. These models

are very similar to several of those developed for ^{137}Cs and described above, the key difference being the need to incorporate a continuous and essentially constant annual fallout input, rather than fallout input over a limited period. The mass balance model for cultivated soils incorporates many of the refinements associated with the comprehensive mass balance model for ^{137}Cs described by the same authors. These include size selective sediment mobilization and deposition and the fate of fresh fallout prior to its incorporation into the plow layer by tillage. The mass balance model has also been extended to incorporate the effects of soil displacement by tillage. Because $^{210}\text{Pb}_{\text{ex}}$ fallout is continuous and essentially constant through time, it is particularly important that the fate of fresh fallout, prior to incorporation into the soil by tillage, should be represented in a physically realistic manner, since this will exert a key control on the $^{210}\text{Pb}_{\text{ex}}$ inventory. If erosion occurs whilst the recent fallout remains at the surface, a relatively small amount of erosion could remove a significant proportion of the fresh input, whereas once the fresh fallout is mixed into the plow layer, the same amount of erosion will remove a much smaller amount of $^{210}\text{Pb}_{\text{ex}}$.

Because the $^{210}\text{Pb}_{\text{ex}}$ fallout is continuous, it is not possible to use a conversion model equivalent to the profile distribution model developed for ^{137}Cs , when investigating soil redistribution on uncultivated soils. That model assumes that the depth distribution is established during the main period of ^{137}Cs fallout and is subsequently truncated by surface erosion. With $^{210}\text{Pb}_{\text{ex}}$, the continuous fallout will replenish the $^{210}\text{Pb}_{\text{ex}}$ removed from the upper part of the soil by erosion and the depth distribution is likely to achieve a quasi steady state. Under these circumstances, a model very similar to the diffusion and migration model developed for ^{137}Cs has been developed.

A different approach to estimating soil erosion rates using $^{210}\text{Pb}_{\text{ex}}$ measurements is described by Wallbrink and Murray [36]. This approach, which was applicable to uncultivated soils combined measurements of both ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ and was recommended for use in areas, such as forests, where local variability of ^{137}Cs fallout made it difficult to establish a ^{137}Cs reference inventory. In this case, attention was directed to the contrasting depth distributions of the two fallout radionuclides, with ^{137}Cs evidencing deeper penetration

than $^{210}\text{Pb}_{\text{ex}}$ in the study area in Tasmania, Australia. The different penetration depths resulted in a change in the ratio between the $^{210}\text{Pb}_{\text{ex}}$ inventory (below a specific depth in the profile) and the inventory of ^{137}Cs below the same depth. The ratio decreased monotonically with depth, down to ca. 8 cm, with values of >2.0 near the surface and <1.0 at depths of ca. 4 cm. The ratio provides a unique measure of the depth within the profile and can be used to estimate the depth of erosion that has occurred at a particular point, either over the period since the occurrence of bomb fallout or since erosion commenced as a result of land clearance or disturbance, if this occurred later. Once the depth distribution of the $^{210}\text{Pb}_{\text{ex}}/^{137}\text{Cs}$ ratio is established for the study site using an undisturbed reference site, bulk cores are collected from individual sampling points and measurement of the ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ inventories permits the ratio to be calculated and the depth of soil eroded to be estimated. The approach is in many respects similar to the use of the profile distribution model to estimate the depth of erosion on an uncultivated soil from measurements of the ^{137}Cs inventory undertaken on bulk cores. In that case, the exponential depth distribution of ^{137}Cs provides a means of estimating the depth of soil removed by erosion, whereas with this approach, the depth distribution of the ratio provides the means of estimating the depth of soil removed from the inventory ratios calculated for individual bulk cores.

Use of ^7Be Measurements In contrast to ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$, the value of ^7Be for documenting soil redistribution rates reflects its short half-life (53 days) and its potential for obtaining information relating to individual events or short periods of heavy rainfall (e.g., 1–2 weeks). However, although there is increasing interest in this potential, relatively little work has been undertaken to date using this fallout radionuclide. One early attempt to use ^7Be to estimate soil redistribution rates employed a mass balance approach, analogous to the repeat monitoring approach developed for use with ^{137}Cs [35]. In this approach, the ^7Be inventory was documented for a number of sampling points immediately prior to the event or short study period. At the end of the event or period of interest, the ^7Be inventory was measured again at adjacent points. The ^7Be fallout input during the intervening period was

estimated using a rainfall collector. If the period studied is sufficiently short to permit the effects of radioactive decay to be ignored, the amount of erosion E (kg m^{-2}) at the sampling point associated with the period between the collection of the two sets of samples can be estimated as:

$$E = \frac{(A_0 + D - A_1)}{C} \quad (6)$$

where A_0 and A_1 are the ^7Be inventories (Bq m^{-2}) at the beginning and end of the study period, respectively; D (Bq m^{-2}) is the fallout input during the study period; and C is the ^7Be content of the eroded soil (Bq kg^{-1}).

The study undertaken by Wilson et al. [35] was undertaken on an erosion plot, and in this case, C could be established by collecting the eroded sediment and measuring its ^7Be content. In that study, C could also have been estimated by using the measurements of soil loss available from the erosion plot to estimate the depth of erosion during the event or study period and to then determine C from the depth distribution of ^7Be established using sectioned cores. In other applications, it is likely to prove more difficult to determine C , which is clearly a critical component of Eq. 6. Use of this mass balance approach must be seen as rather complex, since it involves a set of “before” and “after” measurements, as well as measurements of fallout input during the period of interest and information on the ^7Be content of eroded sediment. As such, it could be seen as providing experimental validation of the potential for using ^7Be to document short-term soil redistribution rates, rather than a more generally applicable procedure.

Most other attempts to use ^7Be for documenting soil redistribution rates (e.g., [32, 33, 106, 107]) have adopted a simpler approach more similar to that employed with ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$. This involves comparison of the inventories documented at a number of sampling points at the end of the period of interest using bulk cores, with a reference inventory, also measured at the end of the period of interest, in order to establish the magnitude of the increase or decrease in inventory in response to soil redistribution. However, the basis of the approach differs in several respects from that associated with ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$. Firstly, as indicated above, attention focuses on short periods, usually a single storm event or a short period of heavy rainfall (e.g., 1–2 weeks). Secondly, study areas are usually

unvegetated. Where a vegetation cover exists, a significant proportion of the ^7Be fallout input to a site is likely to be intercepted and will be retained by the vegetation cover until it is lost by decay. This situation differs from that for ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$, where, by virtue of their much longer half-lives, the fallout will still eventually reach the soil surface after leaf fall or decay of the vegetation cover. This interception of the incident ^7Be fallout by the vegetation cover is likely to cause the fallout input to the soil surface itself to be spatially variable and it is not possible to assume a spatially uniform reference inventory. Equally, if the total inventory contained in both the vegetation canopy and the soil is considered, variable partitioning of the inventory between the two components makes it difficult to link changes in the inventory to soil redistribution. For this reason, the use of ^7Be to study soil erosion and soil redistribution rates is generally restricted to bare cultivated soils.

As emphasized by Figs. 6 and 7, a key feature of the depth distribution of ^7Be in the soil is its restriction to a very shallow layer of the surface soil, generally only ca. 1–2 cm deep. Within this layer, the depth distribution is commonly characterized by a rapid exponential decrease with depth. Little difference is found between the depth distributions associated with cultivated and undisturbed soils, since both are dominated by the fallout input to the surface. Where cultivation occurs, mixing of the ^7Be into the plow layer will cause the activity to fall below the level of detection. The conversion models generally used with ^7Be measurements (e.g., [33, 107]) are very similar to the profile distribution model employed with ^{137}Cs measurements. The ^7Be exponential depth distribution at the reference site can be characterized by the relaxation depth h_0 kg m^{-2} and the mass depth of soil eroded from a sampling point R kg m^{-2} can be estimated by comparing the ^7Be inventory at the sampling point A Bq m^{-2} with that for the reference site A_{ref} Bq m^{-2} , and relating the difference to h_0 , namely,

$$R = h_0 \ln \left(\frac{A_{\text{ref}}}{A} \right) \quad (7)$$

Where the measured ^7Be inventory exceeds the reference inventory, deposition is assumed to have occurred and the mass depth of deposition can be estimated, as with ^{137}Cs (Eq. 4), by dividing the excess ^7Be inventory

by the mean ^7Be content of sediment mobilized by erosion from the upslope contributing area.

The procedure commonly used with ^7Be measurements is to visit the study site after a storm event or short period of heavy rainfall (e.g., 1 week) and to collect cores from a reference site and from the study site. The cores collected from the reference site should include a sectioned core that is used to determine h_0 . By applying Eq. 7, it is possible to estimate the soil redistribution that occurred during the period considered. The short half-life of ^7Be and the emphasis on documenting individual events do, however, impose a number of constraints on the use of ^7Be for estimating soil redistribution which do not apply to the use of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ to estimate longer-term erosion rates covering periods of several decades [14]. Most important is the assumption that the spatial pattern of inventories documented by the cores collected after an event or period of heavy rainfall reflects soil redistribution occurring during that period and that the preexisting ^7Be inventory can be considered to have been spatially uniform across the study area. This means that this approach can only be applied to events or periods of heavy rainfall separated by relatively long periods with little rainfall or at least rainfall of insufficient intensity to cause erosion and soil redistribution. A period of ca. 5 months (i.e., three half-lives) is ideally required to ensure that the spatial variability of the ^7Be inventory inherited from previous events has effectively disappeared as a result of decay. Cultivation of a field provides an alternative means of removing or resetting the signal from past events, since the existing inventory will be mixed into the plow layer and is likely to be below the level of detection. This constraint can seriously limit the potential for using ^7Be measurements to document short-term erosion rates, since effectively only the first period of heavy rainfall following an extended dry period or the cultivation of the study site can be studied. In many environments, the events of most interest may occur later in the wet season or it might be important to consider the whole wet season. As the length of the period of rainfall studied increases, the estimate of erosion or soil redistribution obtained will underestimate the true value to a progressively greater degree. Walling et al. [87] used a hypothetical example to demonstrate that if the period extends to 6–7 weeks and this includes a number of rainfall events

of similar magnitude distributed through that period, the amount of soil eroded could be underestimated by as much as 50%.

Increasing recognition of the potential value of ^7Be in providing a means of documenting short-term erosion rates, to complement the estimates of longer-term erosion rates provided by ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$, has directed attention to the possibility of extending the timescale over which ^7Be measurements can be applied and removing the constraint of only being able to document the first events occurring after a dry period. Such developments will necessarily increase the complexity of the approach and the need for additional information on, for example, fallout inputs, changing values of h_0 , and the distribution of erosion within the period considered. Walling et al. [87] recently reported the development of a novel procedure for applying ^7Be measurements, involving such additional information, which permitted erosion and soil redistribution to be quantified over a period of ca. 3 months. The procedure was developed for application in Chile where soil erosion during the period following forest harvesting was being studied and there was a need to document erosion occurring during the entire wet season.

Estimating Sedimentation Rates and Establishing Chronologies in Depositional Sinks

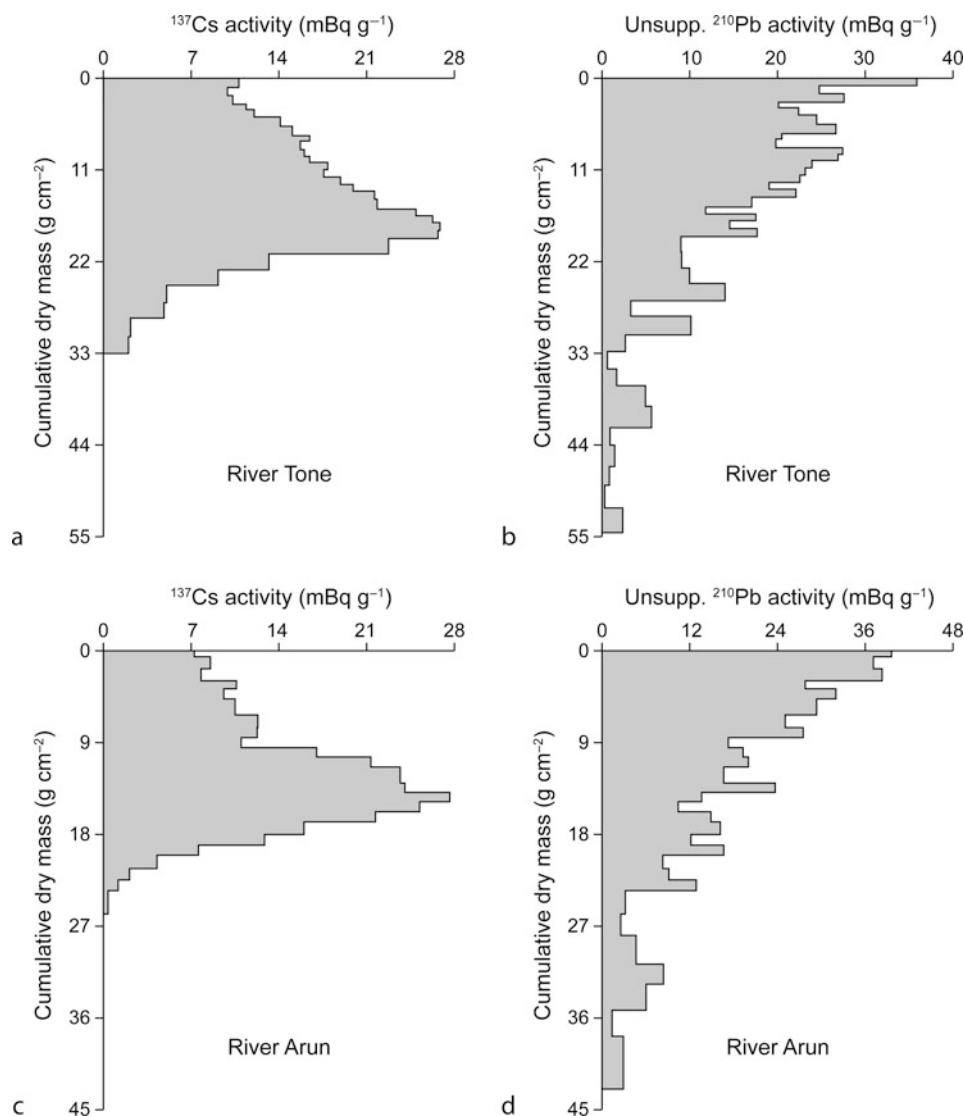
The fallout radionuclides ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ have been widely used to provide estimates of sedimentation rates in depositional sinks, including water bodies such as lakes and reservoirs, wetlands and river floodplains and for establishing the chronology of sediment cores collected from such sinks. In some studies, the emphasis may be on documenting sedimentation rates ($\text{g cm}^{-2} \text{ year}^{-1}$ or mm year^{-1}) and thus quantifying sediment accretion or sediment storage, whereas in others, attention may focus on the chronology. In the latter context, there may, for example, be a need to establish the chronology of changes in the contaminant content of sediment from different depths in a core, in order to reconstruct the pollution history of a study site and the upstream catchment area. Some attempts have been made to use ^7Be for these purposes, but this radionuclide has been less widely applied in sedimentation studies, partly because of its short half-life. With ^{137}Cs , the radionuclide provides information relating

to the period extending from the main period of bomb fallout to the present (i.e., ca. 50 years). In areas where significant Chernobyl fallout occurred, it may be possible to separate the periods before and after 1986. With $^{210}\text{Pb}_{\text{ex}}$, the timescale extends back ca. 100 years. In contrast, where ^7Be is used, it can provide information on very recent sedimentation rates or confirm the presence of recently deposited sediment. Much of the early work using ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ for dating sediment cores and estimating sedimentation rates focused on lakes, and there is now a large body of literature on this topic (e.g., [47–51, 108–110]). Less work has been undertaken on other sediment sinks, such as river floodplains, and these will therefore be emphasized in this contribution, in order to demonstrate the potential for using fallout radionuclides to study the movement of sediment through the landscape.

Use of ^{137}Cs Measurements Figure 10 provides an example of the depth distributions of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ in cores collected from the overbank deposits on the floodplain of two UK rivers, the River Arun and the River Tone, and these provide a useful background for outlining the various approaches that can be used to estimate sedimentation rates and to establish age/depth relationships for sediment cores collected from sediment sinks. The contrasting features of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ fallout, and more particularly the event-related (i.e., bomb testing) nature of the former and the continuous nature of the latter, are reflected by the different principles involved in the use of the two radionuclides. With ^{137}Cs , attention focuses on the known fallout record (see Fig. 2), reflecting the first appearance of significant fallout in the early to mid-1950s and the occurrence of peak fallout in 1963 and in some locations a further short-lived input in 1986, as a result of the Chernobyl disaster. The ^{137}Cs activity of the sediment deposited in a lake, floodplain, or other sediment sink can be expected to reflect this fallout record through two mechanisms. The first mechanism represents direct fallout to surface of the sediment sink that will be incorporated into the depositing sediment. In the case of a lake or reservoir, the ^{137}Cs fallout to the surface of the water body will be dispersed within the water and adsorbed by sediment in the water column, which is subsequently deposited. With a floodplain or marsh, the fallout will be delivered directly to the

surface of the accreting sediment, where it will “label” the surficial sediment. The second mechanism represents a catchment-derived contribution, reflecting ^{137}Cs associated with deposited sediment that was mobilized from the catchment surface, which would itself have received ^{137}Cs fallout. As demonstrated in Fig. 5, the cumulative inventory of the catchment surface, and thus the ^{137}Cs content of eroded sediment, will also vary through time and follow a similar pattern to the direct fallout, peaking around 1963. The precise shape of the ^{137}Cs depth distribution found in a lake or on a river floodplain will reflect the relative contribution of direct fallout and catchment-derived ^{137}Cs to the accumulating sediment [111, 112], but in most situations, the sediment deposited in or around 1963 will possess the highest ^{137}Cs activity. This provides a well-defined time marker. Thus, the clear ^{137}Cs peaks evident at mass depths of ~ 18 and 15 g cm^{-2} for the cores from the Rivers Tone and Arun, respectively, in Fig. 10 can be dated to 1963 and a mean sedimentation rate can be calculated by dividing the total mass or depth of sediment by the period elapsed between 1963 and the date of core collection. Some workers have used the depth of the first appearance of ^{137}Cs as an alternative time marker, which can be dated to ca. 1954. However, this approach is not to be recommended, since there is clear evidence that in most depositional sinks, the “tail” of the ^{137}Cs depth distribution will be influenced by post-depositional downward migration that will cause overestimation of the sedimentation rate. As indicated above, in some locations influenced by Chernobyl fallout, the existence of ^{137}Cs peaks in a sediment core that can be linked to 1963 and 1986 can provide a basis for estimating sedimentation rates for two periods (i.e., 1963–1986 and 1986 to present) [113, 114], and evaluating recent changes in sedimentation rates associated with changes in land use or land management.

The depth at which the ^{137}Cs peak is found is generally considered to be unaffected by downward migration in lake sediment deposits, where mixing is commonly assumed to be of limited importance. However, the possibility of some downward migration must be considered when studying floodplain deposits, since the peak ^{137}Cs activity is commonly found just below the surface in natural soils. The ^{137}Cs depth distribution from a representative undisturbed soil adjacent to



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 10

Depth distributions of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ measured in sediment cores collected from the floodplains of the River Arun and River Tone, UK

the floodplain study site can be used to estimate the likely downward migration rate of the ^{137}Cs peak R_m ($\text{kg m}^{-2} \text{ year}^{-1}$) by dividing the present depth of that peak (expressed as a mass depth) by the time elapsed between the year of peak fallout and the present T_p (years). Thus, if the mass depth at which the peak ^{137}Cs activity occurs in the floodplain core D_{pk} (kg m^{-2}) is known, and R_m has been estimated, the mean sedimentation rate on the floodplain R' ($\text{kg m}^{-2} \text{ year}^{-1}$) can be estimated as:

$$R' = \left(\frac{D_{\text{pk}}}{T_p} \right) - R_m \quad (8)$$

It is important to recognize that R' , as estimated using Eq. 8 or simply from the depth of the ^{137}Cs peak in a sediment deposit, represents an *average* sedimentation rate and that for some sediment sinks, there may be marked inter-annual variability in the amount of deposition. On some floodplains, for example, most deposition may be associated with

infrequent high magnitude floods with recurrence intervals of >10 years.

The above approach to estimating sedimentation rates assumes that the sediment core is from an undisturbed location and therefore that the ^{137}Cs depth distribution reflects the interaction of the fallout input with the progressive sediment accumulation and any associated catchment contribution of ^{137}Cs . In some locations, the floodplain surface may be cultivated and this will cause mixing of the sediment within the plow layer. In this situation, it will no longer be possible to use the ^{137}Cs depth distribution to estimate the sedimentation rate. However, several studies [115–117] have shown that it is still possible to estimate the sedimentation rate, if the floodplain has been regularly cultivated since the beginning of ^{137}Cs fallout. If no sedimentation has occurred, ^{137}Cs will only be found to a depth equivalent to the plow depth. If, however, sedimentation has occurred, its depth can be estimated from the depth to which ^{137}Cs is found below the plow depth, due to the progressive accretion of the surface. Comparison of the total depth of the ^{137}Cs profile which has been homogenized by plowing D_t (kg m^{-2}) with the plow depth D_{pl} (kg m^{-2}) provides a means of estimating the sedimentation rate, namely,

$$R' = \frac{D_t - D_{pl}}{T_p} \quad (9)$$

The need to obtain detailed information on the depth distribution of ^{137}Cs within a sediment core can represent an important constraint in the application of the approach embodied in Eq. 8, which requires information on the depth of the ^{137}Cs peak. Due to likely limitations on the number of core slices that can be assayed, due to cost or time constraints, it may not be possible to document the small-scale spatial variability of sedimentation rates. This may be less of a problem in lakes, where sedimentation is likely to be relative uniform within a small area, but it can be a significant problem for river floodplain investigations, where sedimentation rates can be characterized by marked spatial variability, in response to the floodplain microtopography, distance from the main channel, and the vegetation cover (e.g., [45]). Walling and He [118] provide a means of addressing this problem by describing a procedure for estimating the sedimentation rate from measurements of the total inventory

(Bq m^{-2}) of bulk cores collected from individual points on the floodplain. The bulk cores are assumed to contain the total depth of ^{137}Cs for the sampling points and the key feature of the approach is the estimation of the excess inventory, which is defined as the total inventory less the direct fallout contribution. The latter can be established using cores collected from adjacent reference site located above the level of flood inundation. The excess inventory I_{ex} (Bq m^{-2}) represents the ^{137}Cs input associated with deposited sediment, and its magnitude will reflect the rate of deposition and the ^{137}Cs concentration in the deposited sediment. The latter will clearly vary through time in response to the accumulation of ^{137}Cs in the catchment soils and its subsequent mobilization by erosion. Walling and He [118] describe a model that can be used to estimate the time-weighted mean concentration of the deposited sediment and thus the deposition rate. However, they also propose a simpler and more direct approach for estimating the deposition rate for individual sampling points, based on measurements of the inventories of bulk cores. This makes the assumption that the temporal variation of the ^{137}Cs content of the deposited sediment would have been similar across the floodplain surface and that the main factor accounting for spatial variation in the ^{137}Cs content of deposited sediment will be its grain-size composition. If the sedimentation rate R'_0 ($\text{kg m}^{-2} \text{ year}^{-1}$) and excess ^{137}Cs inventory $I_{ex,0}$ are known for a specific representative point on the floodplain for which a well-defined ^{137}Cs profile obtained from a sectioned core is available, the sedimentation rate R' at other points on the floodplain can be estimated from a comparison of the values of excess inventory I_{ex} and the grain-size composition of the sediment at those points with the equivalent information for the reference point, that is,

$$R' = R'_0 \frac{I_{ex}}{I_{ex,0}} \left[\frac{S_{d,0}}{S_d} \right]^\nu \quad (10)$$

where $S_{d,0}$ and S_d represent the specific surface area of surface sediment ($\text{m}^2 \text{ g}^{-1}$) from the reference point and the sampling point, respectively, and ν = a constant reflecting the form of the relationship between ^{137}Cs concentration and specific surface area, which has been shown to have a value of ca. 0.65–0.7 [84].

It is important that the second term in Eq. 10 should be included, in order to take account of the

influence of grain-size composition on ^{137}Cs activity and the preferential association of ^{137}Cs with finer sediment. In Eq. 10, the assumption is made that the grain-size composition of surface sediment is representative of that for the entire depth profile. The approach involving the use of measurements of the inventory of bulk cores represented by Eq. 10 can also be used for cultivated floodplains, although in this case, R'_0 will be estimated using Eq. 9 rather than Eq. 8. The use of bulk core measurements in the estimation of floodplain accretion rates depends heavily on the assumption that the deposited sediment is characterized by a significant ^{137}Cs activity, which in turn results in readily measurable excess inventories. In some environments, the deposited sediment may contain low levels of ^{137}Cs or no ^{137}Cs , because of the nature of the sediment sources in the upstream catchment or the low inventories existing in the catchment. In this situation, it may be difficult to quantify the limited excess inventory due to the low precision of the gamma spectrometer measurements and the ^{137}Cs depth distribution may represent the only viable approach to quantify the deposition rate. A situation of this nature has been described for the floodplain of a river in central Queensland Australia [119].

Use of $^{210}\text{Pb}_{\text{ex}}$ Measurements Because of the essentially continuous nature of $^{210}\text{Pb}_{\text{ex}}$ fallout, the $^{210}\text{Pb}_{\text{ex}}$ depth distributions associated with sediment sinks do not provide a distinct time marker, as provided by ^{137}Cs . As with ^{137}Cs , the $^{210}\text{Pb}_{\text{ex}}$ content of the accreting sediment will reflect both direct fallout to the lake or floodplain surface and $^{210}\text{Pb}_{\text{ex}}$ associated with sediment mobilized from the upstream catchment. If it is assumed that the sediment surface is progressively buried by deposition, the $^{210}\text{Pb}_{\text{ex}}$ activity of the sediment in a core will decline with depth due to radioactive decay. The rate of decline in the $^{210}\text{Pb}_{\text{ex}}$ activity with depth provides a means of estimating the age of the sediment at different depths and thus the sedimentation rate. The $^{210}\text{Pb}_{\text{ex}}$ depth distributions for the floodplain of the Rivers Tone and Arun shown in Fig. 10, for example, demonstrate that the $^{210}\text{Pb}_{\text{ex}}$ activity declines to zero at a mass depth of about 45–55 kg m^{-2} . Since it is normally assumed that $^{210}\text{Pb}_{\text{ex}}$ activity will decline to near-zero after ca. five half-lives, the sediment at this depth can be estimated to be about 110 years old.

A preliminary estimate of the sediment age at different depths can be obtained by assuming that the $^{210}\text{Pb}_{\text{ex}}$ activity in the deposited sediment declines exponentially with time in accordance with the radioactive decay law, namely,

$$C_d = C_0 e^{-\lambda t} \quad (11)$$

where C_d is the activity at a given time, C_0 is the initial activity of the deposited sediment, λ is the decay constant for ^{210}Pb ($0.03114 \text{ year}^{-1}$), and t is the number of years since burial.

Use of Eq. 11 to estimate sediment age assumes that the activity of the sediment at the time of its initial deposition is constant through time and thus that the activity is not influenced by the sedimentation rate. In the real world, these assumptions may not apply and a range of different models of varying complexity have been developed for interpreting the $^{210}\text{Pb}_{\text{ex}}$ depth distributions in sediment cores, in order to obtain reliable estimates of sediment age or sedimentation rates. In these models, depths are commonly expressed as mass depths (g cm^{-2}) to avoid complications due to sediment compaction and downcore variations in bulk density. The models involve a number of different assumptions. In the simplest model, often referred to as the Constant Flux, Constant Sedimentation rate (CFCs) model, a constant initial concentration for the deposited sediment $C_0 \text{ Bq kg}^{-1}$ and a constant dry mass sedimentation rate $r \text{ g cm}^{-2} \text{ year}^{-1}$, and therefore a constant depositional flux are assumed, resulting in the following relationship:

$$C_m = C_0 e^{-\lambda m/r} \quad (12)$$

where C_0 and C_m are the $^{210}\text{Pb}_{\text{ex}}$ concentrations at the surface of the core and at mass depth m , respectively.

When $\log C_m$ is plotted against m , the resulting plot will be linear and the mean sedimentation rate r can be estimated from the slope of the best-fit line.

In many situations, however, it is unrealistic to assume a constant sedimentation rate, particularly in environments where extreme events play an important role in the operation of erosion and sedimentation processes. Equally, and perhaps more importantly, human impact on erosion rates as a result of land clearance and land use change will mean that sedimentation rates cannot be treated as stationary, but must be seen as being likely to have changed significantly over

the past 100–150 years covered by the $^{210}\text{Pb}_{\text{ex}}$ measurements. Two models, generally referred to as the Constant Rate of Supply (CRS) and Constant Initial Concentration (CIC) models, have been developed to take account of variable sedimentation rates [49, 120, 121]. In the first model, the annual supply of $^{210}\text{Pb}_{\text{ex}}$ to the accreting sediment is assumed to be constant, meaning that the activity of the sediment deposited in a given year will vary according to the amount of sediment deposited. If this is high, the activity will be reduced. With the CIC model, the initial activity of the deposited sediment is assumed to remain constant through time. Neither assumption is necessarily a close fit to reality. In this context, it is important to recognize the two sources of supply of $^{210}\text{Pb}_{\text{ex}}$ to the accreting sediment, namely, direct fallout to the lake or floodplain surface and supply of $^{210}\text{Pb}_{\text{ex}}$ in association with the eroded sediment. Temporal variation of the magnitude of the latter in response to changes in erosion rates and sediment flux will breach the assumption of the CRS model. Equally, variation of the mass of sediment deposited will mean that the initial $^{210}\text{Pb}_{\text{ex}}$ concentration associated with that sediment will vary through time, since the direct fallout input will be mixed with a variable mass of sediment. Furthermore, the initial $^{210}\text{Pb}_{\text{ex}}$ activity of the sediment mobilized from the upstream catchment can be expected to vary as sediment sources change as a result of land clearance and land use change within the catchment. If, for example, gully erosion assumes increasing importance, the $^{210}\text{Pb}_{\text{ex}}$ activity of the sediment mobilized from the upstream catchment might be expected to decline.

Appleby [109] provides an explanation of the procedures involved in applying the CRS model which permits the age t of the sediment at a given depth m to be calculated as:

$$t = \frac{1}{\lambda} \ln \left(\frac{A_0}{A_m} \right) \quad (13)$$

where A_0 and A_m are the total inventories below the surface and below the given depth m , respectively.

The sedimentation rate r g cm $^{-2}$ year $^{-1}$ at time t can also be calculated as:

$$r = \frac{\lambda A_m}{C_m} \quad (14)$$

where C_m is the $^{210}\text{Pb}_{\text{ex}}$ concentration at depth m and time t and A_m is the total inventory below this depth.

With the CIC model, the initial $^{210}\text{Pb}_{\text{ex}}$ concentration of the sediment at depth m and of age t is assumed to have been the same as the current concentration of sediment at the sediment–water interface or floodplain surface C_0 . If the present concentration C of the sediment at this depth is known, its age t can be calculated using the relationship defined in Eq. 11, that is,

$$C = C_0 e^{-\lambda t}$$

The CIC model implicitly assumes that the supply rate of $^{210}\text{Pb}_{\text{ex}}$ is proportional to the sedimentation rate and this situation will not generally apply. For this reason, the CRS model is often assumed to provide more meaningful estimates of the sedimentation rate and, in the case of lakes, this has been confirmed by the similarity of CRS model–derived chronologies to the ^{137}Cs chronology and independent measures of sediment age based on counting of annual laminae. To date, little attention has been given to validating the models when applied to sediment cores from river floodplains.

As with the interpretation of ^{137}Cs depth distributions or profiles, use of the above models to estimate sedimentation rates or sediment age requires detailed information on the $^{210}\text{Pb}_{\text{ex}}$ depth distribution and this may introduce constraints on the number of cores that can be studied. Such limitations may be of limited importance in lakes, where sedimentation rates are unlikely to vary markedly over short distances, but for river floodplains and other terrestrial sediment sinks, such as wetlands, much greater spatial variability may occur and there may be a need to document this. Faced with the need to document the detailed patterns of sediment deposition on river floodplains, He and Walling [122] described a model which could be applied to bulk cores for which a measurement of the total $^{210}\text{Pb}_{\text{ex}}$ inventory was available. In this case, the complexity introduced by the dual source of $^{210}\text{Pb}_{\text{ex}}$, namely, direct fallout to the floodplain surface and $^{210}\text{Pb}_{\text{ex}}$ associated with the sediment mobilized from the upstream catchment, was reduced by focusing only on the latter component. This was calculated by subtracting the local reference inventory A_{ref} from the total inventory A_{total} since the former represents solely the fallout input. The model assumed a constant initial concentration for the deposited sediment C_r and

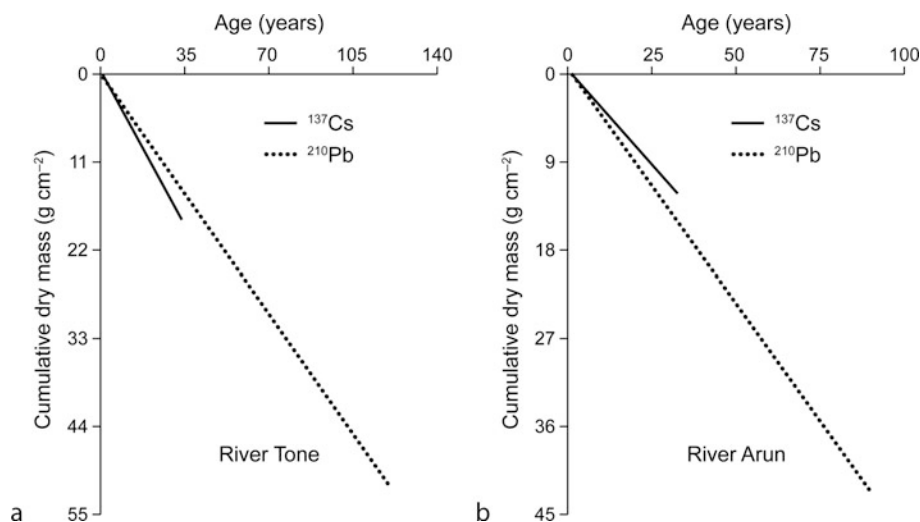
a constant sedimentation rate r (averaged over a period of ca. 100 years) and was termed the CICC model. The mean sedimentation rate for the past ca. 100 years (i.e., ca. five half-lives) can be estimated as:

$$r = \lambda \frac{A_{\text{total}} - A_{\text{ref}}}{C_r} \quad (15)$$

C_r is a key parameter in Eq. 15 and its estimation may prove difficult. It can be determined empirically by using sedimentation traps (e.g., [123]) to collect samples of the sediment deposited during overbank flood events and measuring its $^{210}\text{Pb}_{\text{ex}}$ content. Equally, it can be estimated from measurements of the $^{210}\text{Pb}_{\text{ex}}$ activity of suspended sediment transported during flood events, although potential contrasts between the $^{210}\text{Pb}_{\text{ex}}$ activity of suspended sediment and deposited sediment associated with contrasts in grain-size composition should be taken into account. In the absence of such direct measurements, C_r can also be estimated from measurements of the $^{210}\text{Pb}_{\text{ex}}$ activity of surface sediment from the floodplain and subtracting the direct fallout contribution to that concentration. Figure 11 presents the age–depth relationships derived from the ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ depth distributions presented in Fig. 10 using Eq. 8 for ^{137}Cs and the CICC model for $^{210}\text{Pb}_{\text{ex}}$. There is good agreement

between the estimates provided by the two radionuclides, although the results suggest that deposition rates in more recent years have increased for the River Tone and decreased for the River Arun. Such changes could reflect a range of factors, including changing sediment loads and changes in the frequency and depth of overbank flooding related to either changes in flood generation in the catchment or the effects of progressive floodplain accretion in increasing the bankfull stage. The latter would give rise to a reduction in the frequency and depth of overbank inundation.

In addition to the relatively simple conceptual models for deriving estimates of sedimentation rates within sediment sinks from $^{210}\text{Pb}_{\text{ex}}$ depth distributions, described above, more complex inductive models that permit the interpretation of $^{210}\text{Pb}_{\text{ex}}$ profiles in situations where both sedimentation rates and $^{210}\text{Pb}_{\text{ex}}$ fluxes vary with time have been developed [51, 124]. In this case, signal theory is used to compare the actual $^{210}\text{Pb}_{\text{ex}}$ profile with model profiles produced using realistic values of the various parameters and, by minimizing an objective function, it is possible to deduce the model form and the sedimentation rate. The Sediment Isotope Tomography (SITS) model described by Carroll and Lerche [51] is probably the



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 11

A comparison of the age–depth relationships established for the floodplain cores illustrated in Fig. 10, using the ^{137}Cs depth distribution and the $^{210}\text{Pb}_{\text{ex}}$ depth distribution coupled with the CICC model

most widely applied model of this type, and some success has been achieved in incorporating bioturbation and mixing into the model structure.

For most sediment sinks, it is reasonable to assume quasi-continuous sedimentation, even though sedimentation rates may vary from year to year and over longer periods. This is the situation on many river floodplains where the river inundates the floodplain several times a year and accretion is ongoing. In some situations, however, sedimentation may be episodic and this introduces complications for interpreting the sediment profile, as this may reflect episodic accumulation of sediment, separated by periods of several years, or even a decade, with no sedimentation. A situation of this nature has been described for the floodplains of the Beni and Mamore Rivers in the Andean-Amazonian Foreland [125], but it still proved possible to use $^{210}\text{Pb}_{\text{ex}}$ measurements to reconstruct the sedimentation history of the floodplain over the past 100 years and to demonstrate the importance of extreme ENSO (El Niño/La Niña) events within this history. Eleven major events in the past 90 years were identified. In this case, a model referred to as the CIRCAUS (constant initial reach clay activity, unknown sedimentation) model was used. When applying this model, the $^{210}\text{Pb}_{\text{ex}}$ activity is normalized to reflect the clay ($<4\ \mu\text{m}$) content of the sample, since most of that activity was shown to be associated with the clay fraction. The same CIRCAUS model was successfully used to study the chronology of floodplain sedimentation for the Strickland River in Papua, New Guinea [126].

Using ^7Be Measurements In comparison with ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$, relatively little use has been made of ^7Be to date for investigating sedimentation rates within sediment sinks in fluvial systems. Its use for this purpose has, however, been more fully exploited in coastal environments. Because of the short half-life of ^7Be (53 days), it can provide an effective means of identifying recently deposited sediment. Older sediment can be assumed to contain no ^7Be . This approach has been successfully used in shallow marine environments [127, 128] and a similar approach was employed in a study of the deposition and resuspension of PCB-rich sediments in the Fox River, Wisconsin, USA [129]. That study emphasized that appreciable short-term deposition and remobilization occurred, with short-term

accumulation rates greatly exceeding the longer-term rates documented using ^{137}Cs measurements.

Although it is normally assumed that fallout radionuclides such as ^7Be will be preferentially associated with fine sediment, and particularly the clay fraction, recent work has demonstrated how coarser sand-sized ($63\ \mu\text{m}$ – $2\ \text{mm}$) fluvial sediment can also retain ^7Be and that the radionuclide can be used to provide information on the residence time of such sediment in channel bars in rivers and on the sedimentation rates associated with such bars [130]. In a study of the Ducktrap River in Maine, USA, the evolution of channel bars associated with obstructions caused by large woody debris and the associated bar accumulation rates were investigated using ^7Be measurements and the application of a variant of the CIC model to interpret the depth distribution of ^7Be in the channel bars.

^7Be measurements can also be used in a similar manner to ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ to estimate floodplain sedimentation rates. However, because of its short half-life, attention focuses on the sedimentation associated with individual events, rather than longer-term average rates of accretion. If shallow cores are collected from a floodplain after a flood event, the ^7Be inventories of those cores will reflect both the recent fallout to the floodplain surface and ^7Be associated with the sediment mobilized from the upstream catchment and deposited on the floodplain. If the local reference inventory for the time of sampling $A_{\text{ref}}\ \text{Bq m}^{-2}$ is established, by collecting shallow cores from an adjacent reference site above the level of the inundating floodwaters, subtraction of this value from the inventories measured for the individual sampling points enables the excess inventory $A_{\text{ex}}\ \text{Bq m}^{-2}$ associated with the deposited sediment to be calculated. The magnitude of this value will directly reflect the amount of sediment deposited and if the ^7Be concentration of the newly deposited sediment $C_d\ \text{Bq kg}^{-1}$ is known, the amount of sediment deposited $D_f\ \text{kg m}^{-2}$ can be estimated, namely,

$$D_f = \frac{A_{\text{ex}}}{C_d} \quad (16)$$

C_d is an important parameter in Eq. 16 and can prove difficult to establish. It could be measured directly by collecting samples of the deposited sediment using a sedimentation trap, or the ^7Be activity of the

suspended sediment transported by the river sampled during the flood event could be used as a surrogate. In both cases, however, the effects of the particle size composition of the sediment in influencing its ^7Be concentration must be considered. If the first approach is used, it is unlikely to prove feasible to install sediment traps adjacent to each sampling point and only a single sample or a composite sample comprising sediment collected from several traps may be available. In this situation, the ^7Be activity of the sample of deposited sediment C_t must be adjusted to provide an estimate of C_d for each sampling point by taking account of any difference in particle size between the sediment deposited at the sampling point and the sediment collected using the sediment trap. As with Eq. 10, which applied a similar approach for correcting ^{137}Cs activity to reflect the variation of the grain-size composition of sediment between individual sampling points, the measure of specific surface area ($\text{m}^2 \text{g}^{-1}$) provides a convenient means of comparing the particle size composition of samples. If surface samples are collected adjacent to each sampling point and the specific surface area of those samples S_d and the sample collected by the sediment trap S_t is measured, C_d for individual sampling points can be estimated as:

$$C_d = C_t \left[\frac{S_d}{S_t} \right]^\nu \quad (17)$$

where ν = a constant reflecting the form of the relationship between ^7Be concentration and specific surface area with a value of ca. 0.64 [53]. If the estimate of C_d is based on measurement of the ^7Be activity of suspended sediment, the values of C_t and S_t in Eq. 17 must be replaced by the values for the suspended sediment.

The procedure outlined above has been successfully used to document the spatial variability of sediment deposition during a single overbank flood event within a reach of the River Culm floodplain in Devon, UK [53]. When applying this approach, it is important to ensure that the excess inventory inherited from previous events is minimal and this means that the overbank flood studied should be preceded by a period of ca. 5 months with no inundation of the floodplain. To overcome this constraint, it might be possible to collect cores from the floodplain both before and after the event, but this may introduce problems in terms of

forecasting the occurrence of the flood that is to be studied and collecting samples from closely adjacent points, so that the pre- and post-flood ^7Be inventories can be directly compared. Since most floodplain surfaces are vegetated, the presence of grass and other vegetation can also introduce problems, but this is less significant than for studies of soil redistribution within fields. The shallow cores should include the vegetation mat, and because the amount of sedimentation is calculated from the excess inventory, no assumption is made as to whether the deposited sediment is held within the vegetation mat or if it is on the floodplain surface.

As indicated above, most work on using fallout radionuclides to estimate sedimentation rates and sediment chronologies for sediment sinks has focused on lakes and reservoirs, which commonly represent simpler systems than other sinks, such as river floodplains and wetlands. For example, it is generally assumed that the fallout input to a lake surface will be mixed within the body of the lake and will be attached to depositing sediment particles, which may already contain the same radionuclide by virtue of having been mobilized from the upstream catchment. As such, the two components of the total activity of the deposited sediment associated with a given year will be closely integrated and contained within the sediment deposit for that year. In the case of a river floodplain, however, the fallout to the floodplain surface will be deposited on the sediment surface over the course of the year and will be distributed within the shallow surface layer according to its initial distribution. If overbank sedimentation occurs during a flood, this will be deposited on top of the fallout input and the two may not be integrated. Furthermore, depending on the relative magnitude of the deposition rate and the relaxation depth of the initial distribution, the fallout input may be distributed through the sediment deposited over several years, or alternatively, it may only penetrate the upper part of the sediment deposited in a given year. Equally, there is likely to be greater temporal variability in the grain size composition of sediment deposited on a river floodplain, according to the magnitude of the flood events and this may influence the depth distribution of the fallout radionuclide, with coarser horizons being marked by lower activity. As noted above, Aalto and his coworkers have overcome this problem in their

studies of the floodplains of large rivers, such as the Amazon and the Strickland River in Papua New Guinea [125, 126], by normalizing the $^{210}\text{Pb}_{\text{ex}}$ measurements for the clay ($<4\ \mu\text{m}$) content of the individual sediment sections recovered from the cores.

Tracing Suspended Sediment Sources and Sediment Transfer

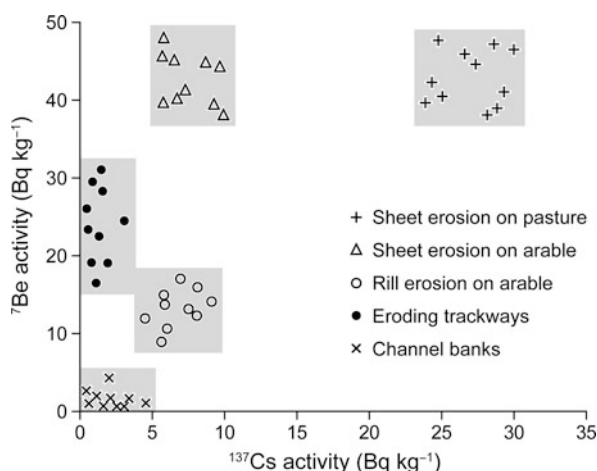
In addition to their use for estimating soil redistribution rates and rates of sedimentation in sediment sinks, such as lakes and river floodplains, fallout radionuclides also provide important opportunities for tracing suspended sediment sources and for developing an improved understanding of sediment transfers in catchments and river basins. Concern for the off-site impacts of soil erosion in providing increased inputs of fine sediment to watercourses and for the many adverse impacts of increased fine sediment loads in river systems has focused attention on the need for sediment management or control strategies aimed at reducing sediment mobilization and transport. A range of control measures have been developed and successfully employed to reduce fine sediment fluxes, but the design of effective control strategies requires that control measures should target the key sediment sources and transfer pathways, in order to optimize the return on expenditure. Expenditure on reducing soil loss from cultivated areas could, for example, bring limited return if much of the sediment transported by a stream is mobilized by channel erosion. For this reason, an assessment of the primary suspended sediment sources in a catchment is often seen as an essential precursor to designing and implementing a sediment control strategy [131].

Sediment source fingerprinting techniques are increasingly used to provide information on sediment sources [132]. In essence, these techniques involve collecting samples of the fine sediment transported by a stream or river or deposited on its bed and comparing the chemical and physical properties of this sediment with those of potential sources. Such comparisons can be used to identify the primary source of the sediment and if a numerical mixing model is used (e.g., [131–135]), it is possible to quantify the relative contribution of a range of potential sources. In this context, *source* can represent *spatial sources* (e.g., different

subcatchments or zones within a catchment), but more often emphasis will be placed on what can be referred to as *source types*, which reflect the nature of the erosion process or sediment source, for example, rill and interrill erosion on cultivated fields, and on land occupied by pasture or rangeland, gully erosion, erosion of tracks and unmetalled roads, and bank and channel erosion.

The success of sediment source fingerprinting studies depends heavily on the availability of tracer or fingerprint properties that are able to discriminate between potential sources. If the aim is to assess the relative importance of a number of different sources, several fingerprint properties will commonly be used. A wide range of sediment properties have been successfully used as source fingerprints, and these include inorganic and organic geochemistry, mineral magnetism, sediment color, stable isotopes and radionuclides. Fallout radionuclides, and particularly ^{137}Cs , ^7Be , and $^{210}\text{Pb}_{\text{ex}}$, have proved particularly useful as source fingerprints, since their fallout origin makes them essentially independent of the underlying geology or soil type and underpins their successful use for discriminating between different potential sediment source types [135]. In this context, the surface of a catchment will be exposed to direct fallout and the activity found in surface soils will be substantially greater than that associated with sediment exposed in channel banks and ditches or in gullies. Equally, where the soil is cultivated, the activity of ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ in the surface soil can be expected to be significantly lower than in an undisturbed soil, such as might be found in an area of pasture or rangeland. Furthermore, the different nature and behavior of ^7Be fallout can provide potential for further discrimination, since this will only label the upper 1 or 2 cm of a cultivated soil, whereas ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ will be mixed throughout the plow layer. This can provide a means of distinguishing sediment mobilized by sheet or interrill erosion and rill erosion, since the surface layer mobilized by sheet or interrill erosion will be characterized by high ^7Be activity, whereas the ^7Be activity of sediment mobilized by rill erosion will be lower, as it includes sediment mobilized from a greater depth within the plow layer.

Figure 12 provides a further demonstration of the ability of fallout radionuclides, in this case ^{137}Cs and ^7Be , to discriminate between potential sediment



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 12

The use of ^{7}Be and ^{137}Cs activity to discriminate between five potential suspended sediment sources in a small 1 km² catchment in Devon, UK

sources in a small (ca. 1 km²) catchment in Devon, UK. By considering both the ^{137}Cs and ^{7}Be activities, it is possible to distinguish sediment mobilized from five potential sources. Sediment mobilized by sheet erosion from cultivated and pasture fields is characterized by a high ^{7}Be activity, by virtue of its surface origin, but the cultivated and pasture sources are clearly distinguished by their ^{137}Cs activity, which is much lower in the cultivated soil, due to the homogenization of the plow layer by tillage. Where rill erosion is the dominant erosion process on cultivated land, the ^{137}Cs activity of the mobilized sediment is close to that associated with sediment mobilized from cultivated land by interrill erosion, but the ^{7}Be activity is much lower, due to the inclusion of soil from below the surface containing little or no ^{7}Be . In contrast, channel banks are characterized by both low ^{137}Cs and ^{7}Be activities, because of the limited exposure of the face to fallout and the presence of both radionuclides in only the upper part of the bank profile. Eroding trackways are also distinguished by their radionuclide fingerprint. As erosion proceeds, the existing ^{137}Cs inventory is removed and the eroded surface material is characterized by a low ^{137}Cs activity. However, the ^{7}Be activity of sediment from this source is seen to be intermediate between that associated with channel banks and sediment mobilized

from the surface of cultivated and pasture fields by sheet erosion. Because the ^{7}Be content of surface material is continually replenished by fresh fallout, the surface of eroding trackways can be characterized by an appreciable ^{7}Be activity. The sediment mobilized from the surface of eroding trackways is, however, likely to have a higher activity than that mobilized by rill erosion from cultivated fields, since the rills will penetrate deeper and mobilize some sediment with a low ^{7}Be activity. The addition of $^{210}\text{Pb}_{\text{ex}}$ to the results presented in Fig. 12 is likely to demonstrate further contrasts between the potential sources, since the ongoing nature of $^{210}\text{Pb}_{\text{ex}}$ fallout will provide some replenishment of the stock of this radionuclide in the surface horizons of eroding areas, whereas the effective cessation of ^{137}Cs fallout in the late 1970s means that no replenishment has occurred since that time.

The discrimination between various potential sediment sources provided by ^{7}Be and ^{137}Cs activity that is demonstrated in Fig. 12 means that these two fallout radionuclides are valuable fingerprint properties and the potential to distinguish sources could be increased further by looking also at $^{210}\text{Pb}_{\text{ex}}$ activity. However, additional fingerprint properties are likely to be required if a mixing model is used to provide definitive estimates of the relative contribution of the potential sources to the sediment sampled at the outlet of a catchment, in order to avoid problems of equifinality. Statistical procedures, such as multiple discriminant function analysis [131, 136], can be used to identify the optimum set of fingerprint properties or composite fingerprint, as it is often termed, for discriminating between all of the potential sources. There is also a need to recognize that contrasts in particle size composition and organic content between the sediment sampled at the catchment outlet or in the river, and the source material can mean that it is not possible to compare sediment and source material properties directly, and that some correction of the property values for particle size and organic matter content may be required [136]. Recent work has recognized the many uncertainties associated with characterizing the fingerprints of potential sources, and Monte Carlo techniques and Bayesian statistics have been used to explicitly recognize this uncertainty within the mixing models used to establish the relative contribution of different sources [137, 138].

The use of fallout radionuclides as sediment source fingerprints, both alone and with other sediment properties, has been successfully demonstrated at a range of scales. At the plot scale, for example, the contrasting depth distributions of ^7Be , ^{137}Cs , and $^{210}\text{Pb}_{\text{ex}}$ have been shown to provide a means of establishing the relative contribution of sheet and rill erosion to the sediment output from small erosion plots under rainfall simulators and documenting the temporal variation of the sediment contributions from the two sources [139, 140]. At the larger field scale, Whiting and coworkers [141] investigated the erosion processes operating on a 6.03 ha cultivated field near Treynor, Iowa, USA, and used information on the different depth distributions associated with ^7Be , ^{137}Cs , and $^{210}\text{Pb}_{\text{ex}}$ and measurements of the radionuclide content of the sediment output from the field during a storm event in mass balance calculations to estimate the areal extent of rill and sheet erosion and the characteristic depths of erosion associated with each process. Rill erosion was found to produce 29 times more sediment than sheet erosion, with rills eroding 0.38% of the field to a depth of 35 mm and sheet wash eroding 37% of the field to a depth of 0.012 mm. Within larger catchments, He and Owens [142] describe the use of ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$, and ^{226}Ra activities to establish radionuclide fingerprints for cultivated land, uncultivated land, and channel banks within the 273 km² catchment of the River Culm in Devon UK and the application of these fingerprints to estimate the relative contribution of the three sources to the sediment output from the catchment. Similarly, Whiting et al. [143] working within the Yellowstone River catchment in Montana, USA, used measurements of ^7Be and $^{210}\text{Pb}_{\text{ex}}$ to discriminate two primary sediment sources, namely, the catchment surface and eroding channel banks, and to estimate the changing relative importance of these sources along a 423 km² reach of the river.

In addition to their use as source fingerprints, fallout radionuclides have also been used as tracers to provide an improved understanding of the movement or transfer of sediment through river systems. In the study undertaken in the Yellowstone River in Montana [143] referred to above, radionuclide activities in the suspended sediment sampled at different locations along the river were used to investigate the sediment transport dynamics. Fine sediment transport distances

were estimated from the exponential decrease in ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$, and ^7Be concentrations downstream, and the transport distance was found to increase from a few km in the headwaters to hundreds of km downstream. Measurements of the ^{137}Cs and ^7Be activities of suspended sediment and fine bed sediment collected within the catchment of the River Clyst in Devon, UK were also used by Blake and coworkers [53] to explore the importance of short-term storage and remobilization of sediment within the channel system to downstream sediment delivery. The ratio of ^7Be to $^{210}\text{Pb}_{\text{ex}}$ has also been successfully employed to assess the importance of recently resuspended bottom sediment to the suspended sediment moving within the Savannah Estuary, USA [144]. In this estuary, ^7Be activities were very low in the bottom sediment, when compared to those associated with recent suspended sediment inputs, while there was little difference in $^{210}\text{Pb}_{\text{ex}}$ activity between the two sediment types.

As noted above, ^7Be has been found to provide a useful chronometer for coarser sand-sized fluvial sediment (transitional bedload) deposited in channel bars [130] and the labeling of this coarser sediment with ^7Be has also been exploited to document bedload transport velocities and the residence time of bed deposits [145]. In that study undertaken in Vermont, USA, the ratio of ^7Be to total ^{210}Pb was used in preference to ^7Be alone, since use of the ratio takes account of variations in grain size between samples. When sediment is deposited and stored on the riverbed, its ^7Be activity declines through time through decay, whereas the ^{210}Pb activity remains essentially constant. The magnitude of the ratio can therefore provide information on the short-term history of sediment deposition, remobilization, and replenishment at a site. The existence of a flood control dam on a river also provided the opportunity to investigate bedload transport velocities. Closure of the dam gate in early December caused the trapping of bedload and when the dam gate was opened in late March, the bedload which had been stored for several months and therefore had a near-zero ^7Be activity was released. By studying the downstream migration of this “dead” sediment, it was possible to document bedload transport velocities and to establish the length of the reach over which the influence of the dam in controlling bedload transport was evident.

Field and Laboratory Techniques

A detailed description and discussion of the field and laboratory techniques and procedures associated with the use of fallout radionuclides in the study of erosion and sedimentation falls outside the scope of this contribution. The reader is referred to the *Handbook for the Assessment of Soil Erosion and Sedimentation Using Environmental Radionuclides* edited by F. Zapata and published by Kluwer Academic Publishers in 2002 [29] for a detailed discussion of most of the field techniques and laboratory procedures employed when using ^{137}Cs measurements. Many of these are common to ^7Be and $^{210}\text{Pb}_{\text{ex}}$. Some further description and discussion of techniques will also be included in the case studies or examples presented at the end of this contribution. However, a brief overview of some key features of the field and laboratory techniques employed and important issues that frequently need to be addressed is provided below. This deals firstly with field techniques for sample collection, and secondly with laboratory analysis of the samples collected.

Field Techniques for Sample Collection

In most cases, the application of fallout radionuclides in soil erosion and sedimentation investigations involves the collection of samples from the field, for example, soil cores for use in studies of soil redistribution, cores from lakes or river floodplains for documenting sedimentation rates or establishing the chronology of sediment deposits, or samples from potential sources and samples of suspended sediment for use in sediment source fingerprinting investigations. As with most environmental sampling, there is an important need to ensure that the samples are representative and that they will provide results that can be readily interpreted. A carefully designed sampling program is therefore an essential precursor to the program of laboratory analysis. The need for representative samples involves both a temporal and a spatial dimension and both dimensions are important. When collecting samples of suspended sediment for establishing suspended sediment sources, for example, the temporal dimension can be very important and it is necessary to recognize that the properties of the sediment passing a sampling point may vary considerably through time, in response to different stages of the hydrograph and

the routing of sediment from different parts of the catchment which may contribute at different times. It will therefore often be necessary to collect time-integrated samples, in order to provide representative information on the source of the sediment. Time integrating suspended sediment samplers or traps [146] provide one means of collecting such samples. These traps also have the advantage that they collect relatively large samples. Sizeable amounts of sediment may be required for subsequent laboratory analysis and if instantaneous samples are collected, a continuous flow centrifuge may be required to recover the sediment from large water samples [147].

When sampling fields to investigate patterns and rates of soil redistribution or collecting cores from lakes or river floodplains to document sedimentation rates, the temporal dimension will generally be of very limited importance for studies employing ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ since these radionuclides provide information on medium- or longer-term process rates (i.e., 50–100 years) and their inventories are unlikely to be influenced by the conditions immediately preceding the time of sampling. However, when ^7Be is being used, the timing of sampling can be a key consideration in terms of providing definitive results, since there is need to avoid the possibility of past events influencing the measurements obtained.

Important issues related to the spatial dimension extend across the sampling associated with most applications of fallout radionuclides in the study of erosion and sedimentation. A key consideration here is that the cores, which frequently represent the main means of sample collection in studies of soil redistribution or sedimentation, are generally of relatively small diameter and are therefore characterized by a small or, perhaps more correctly, a very small, surface area. Sampling points must be carefully selected and the sampling scheme must be carefully designed to obtain representative data. When studying soil redistribution within fields, a density of 20 cores ha^{-1} is likely to be considered fairly dense, but such cores are typically ca. 8 cm in diameter and the cores will therefore typically only cover <0.001% of the surface area under investigation. Transects and sampling grids are frequently employed in studies of soil redistribution or sedimentation patterns and rates to provide representative data, but it is important to also recognize the uncertainty

that is likely to be introduced by microscale sampling variability, such that if the core was taken a few centimeters away from the initial sampling point, a different inventory value might be obtained [148]. Problems of sampling variability assume particular importance when samples are collected to define the local reference inventory [149, 150]. In addition to carefully selecting a stable undisturbed site that should provide a reliable estimate of the reference inventory, there is a need to collect an appreciable number of cores from the site, the results for which can be averaged, in order to obtain a reliable estimate of that inventory. The inventories measured for individual cores collected from a reference site can frequently be characterized by a coefficient of variation of ca. 20%, or perhaps more, and this means that about 16 cores would need to be collected to provide a mean value that was within 10% of the true mean at the 95% level of confidence. The use of a scraper plate [151], which collects the sample from a larger area, as an alternative to cores, can help provide more representative samples. Similar considerations apply when collecting samples to be used to represent the fingerprint properties of potential sources in source fingerprinting investigations. It is important to ensure that the samples collected are representative of the sediment that might be mobilized from a given source and the uncertainty introduced by representing the final fingerprint by mean property values must be recognized. Problems of spatial variability are generally less important when selecting the locations for collection of cores to be used for determination of the chronology of sediment deposits, but again it is important to collect the core from a representative position and it is generally considerably more difficult to appreciate potential subsurface variability or heterogeneity than variability related to readily visible surface conditions.

Corers will be a key requirement in many investigations employing fallout radionuclides to document soil and sediment redistribution and sedimentation. A range of different corers have been developed for use in different materials and environments, including lakes and reservoirs, wetlands, and soils and these are described in the literature. Sample size may be an important constraint, particularly with sectioned cores, and large diameter corers may be required to provide sufficient mass for the analysis of radionuclide activity, if the core is sectioned into small depth

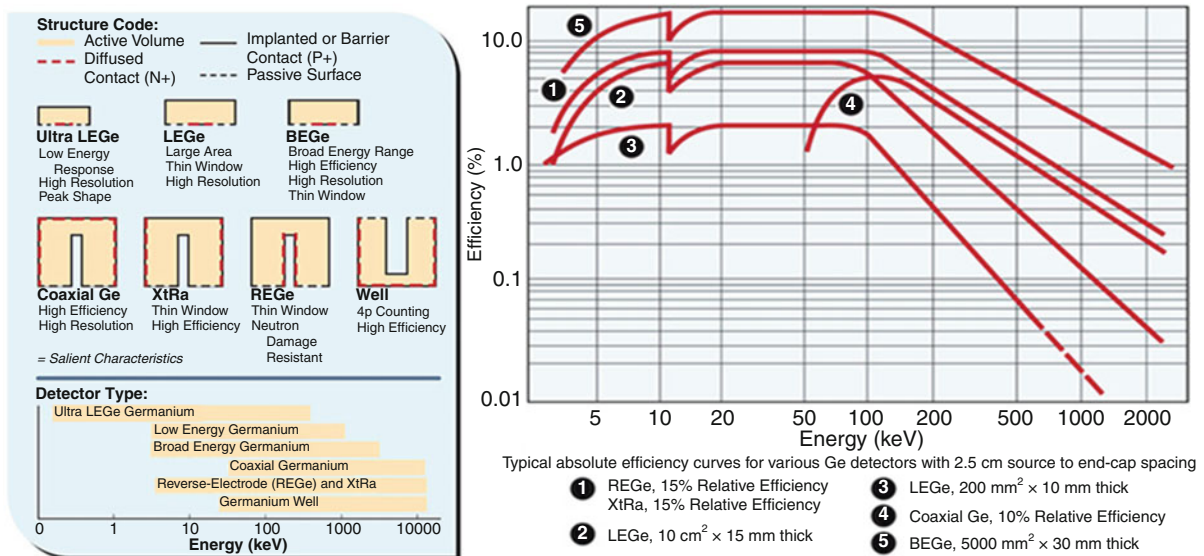
increments (e.g., 1 cm). For soils, cylindrical core tubes are commonly used, but if soils are compacted and dry, it may prove difficult to propel the core tube into the soil to the required depth (e.g., 60 cm). In this situation, a motorized percussion driver is often used to insert the core tube and a winch device may be required to extract the core tube. A scraper plate [151] can provide an alternative to a core, particularly if detailed depth incremental sampling is required, but use of these devices can prove very time consuming relative to the use of sectioned cores. One particular problem commonly faced when using ^7Be measurements is the need to section a shallow core into very fine depth increments (e.g., 2 mm), in order to define the precise shape of the depth distribution, which may only extend down to ca. 15 mm. In order to obtain samples of sufficient mass for the individual depth increments, a corer with a sizeable diameter will be required and devices involving a piston moved by a screw thread have been developed to permit the extrusion of the core to be carefully controlled.

Laboratory Procedures and Techniques

A key feature of the use of fallout radionuclides is that for most applications, the activities of the radionuclides of interest can be readily measured by gamma spectrometry and, provided suitable detectors are available, the radionuclides can be measured contemporaneously during the same counting operation. This removes the need for separate analyses. Gamma spectrometry is therefore the main analytical technique employed in fallout radionuclide investigations, although, in some studies focusing on ^{210}Pb , alternative analytical procedures using radiochemical separations and alpha counting are used [152]. However, this approach involves chemical extractions which require a substantial time input; furthermore, it can only provide measurements of total ^{210}Pb . Additional gamma spectrometry measurements would still be needed to quantify the ^{226}Ra and supported ^{210}Pb activity and thus calculate the $^{210}\text{Pb}_{\text{ex}}$ activity. In the case of sediment cores, collection of samples from depths below the likely influence of $^{210}\text{Pb}_{\text{ex}}$ fallout, where the measured ^{210}Pb activity can be assumed to represent supported ^{210}Pb , can provide a means of establishing the likely supported ^{210}Pb activity of the sediment,

although this value could change upcore. Another important advantage of gamma spectrometry is that only minimal sample preparation is required. Samples need only to be dried and disaggregated and possibly sieved prior to placing in containers to be counted. However, one key limitation of the assay of environmental samples by gamma spectrometry is the long count times involved. These are commonly of the order of 6–12 h per sample, and where activities are lower, count times of 24 h or more may be necessary. The long count times will frequently impose important limitations on the number of samples that can be analyzed within an investigation, since gamma detectors are relatively expensive and it is likely to be possible to operate only a limited number of detectors in a laboratory. With ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$, dealing with large backlogs of samples may not represent too great a problem, apart from the time required to complete a study. However, with ^7Be , the short half-life means that samples must be analyzed relatively quickly after collection and this will commonly impose a limit on the number of samples that can be collected and analyzed within a particular study. It is important that such studies should be carefully planned to take account of these constraints to ensure that the final data set is complete.

Further details of the instrumentation required for analysis of samples by gamma spectrometry and of the procedures employed when analyzing soil and sediment samples are available in specialist handbooks [153]. Measurement of ^{137}Cs , ^7Be , and $^{210}\text{Pb}_{\text{ex}}$ activities is undertaken using Germanium detectors, which have a high resolution and are therefore able to distinguish the signals produced by different radionuclides at similar energy levels. Figure 13 provides a useful summary of the main features of the various Germanium gamma detectors that are available, and attention should be directed to both the geometry and the energy ranges of these detectors. Although ^{137}Cs and ^7Be can be measured on standard gamma detectors, the measurement of ^{210}Pb requires a low-energy detector that is capable of undertaking measurements at the low-energy level (46 keV) associated with this radionuclide. Fortunately, the low-energy detectors needed for ^{210}Pb can, if required, also be used to measure ^{137}Cs and ^7Be , as well as to make the necessary measurements to determine the ^{226}Ra content of the sample required to calculate the supported ^{210}Pb activity. Where large samples (e.g., 1 kg) are involved, Marinelli beakers are commonly used as the sample container, as these fit over the detector head and are designed to optimize the positioning of the sample relative to the detector for



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 13

The configuration, geometry, and response characteristics of Germanium detectors used for gamma spectrometry [154]

counting. Where small samples (e.g., 50–100 g) are counted, these can be put into small plastic pots that are placed on the detector head. For smaller samples (e.g., ca. 10 g), such as may be obtained from the sectioning of small diameter sectioned cores or from small samples of suspended sediment collected from a river, a well detector is commonly used. In this case, the sample is placed in a tube which is inserted into a hole in the top of the detector in order to optimize the counting geometry.

Although gamma spectrometry is relatively undemanding in terms of sample preparation, the detectors must be carefully calibrated to provide accurate measurements [153]. In the case of ^{210}Pb measurements, self-absorption effects can introduce problems and must be considered. The principle of measurement, which depends on the interaction of gamma rays (photons) produced by the disintegration of the radionuclide, an essentially random process, with the germanium crystal, means that the precision of the results obtained is heavily dependent on the number of disintegrations counted. This in turn depends on the activity of the sample, the efficiency of the detector, and the counting time. All results should therefore be accompanied by an indication of their likely precision, and this uncertainty must be recognized when using the results. The precision is typically of the order of $\pm 5\text{--}10\%$ at the 95% level of confidence. In the case of ^{137}Cs and ^7Be , the radionuclides can be measured directly by gamma spectrometry, but determination of $^{210}\text{Pb}_{\text{ex}}$ requires two measurements. The first quantifies the total ^{210}Pb content of the sample and the second establishes the ^{226}Ra content, which is used to calculate the supported ^{210}Pb activity by assuming equilibrium between ^{226}Ra and ^{210}Pb . The $^{210}\text{Pb}_{\text{ex}}$ activity is then calculated by subtracting the supported ^{210}Pb activity from the total ^{210}Pb activity. The Radium-226 content of a sample is usually determined by measuring the activity of its daughters ^{214}Bi and ^{214}Pb , because its direct measurement is complicated by interference from other radionuclides. The need for two measurements and the subsequent subtraction means that the precision of the final estimates of $^{210}\text{Pb}_{\text{ex}}$ activity is usually lower than that associated with ^{137}Cs and ^7Be measurements.

Although gamma spectrometry measurements are generally undertaken in the laboratory, to which

samples collected in the field are transported, attention has been directed to the potential for using in situ detectors [155, 156]. In this case, the detector is deployed in the field in a portable configuration. The detector head is positioned ca. 1.5 m above the soil surface, and the gamma rays received from the soil are counted. Because gamma rays will be received from a relatively large area, the approach overcomes some of the sampling variability problems associated with the collection of small core samples and this, as well as the possibility of obtaining immediate results, are seen as important potential advantages. However, problems can arise in calibrating the measurements, since the precise area of soil from which gamma rays are received can prove difficult to define, due to attenuation effects. Equally a given radionuclide inventory can produce different signals according to the depth distribution of the radionuclide in the soil, and there is a need to make certain assumptions regarding the likely form of the depth distribution. Self-absorption effects also mean that the approach is unlikely to be applicable to ^{210}Pb measurements. Furthermore, in most environments, substantial count times (e.g., ca. 1 h) will be required and this may limit the number of points for which measurements can be collected. As a result, the use of field portable detectors has not been widely adopted and further research is required to assess more fully the potential and limitations of the approach. However, the approach has proved more successful for ^{137}Cs measurements in areas that received substantial amounts of Chernobyl fallout, since the high activities in such areas greatly reduced the count times required [157]. Furthermore, with such high activities, the ^{137}Cs signal is more readily distinguished from those of other radionuclides with similar energy levels and low cost scintillation detectors can be used in place of more costly Germanium detectors.

Uncertainty Considerations in the Use of Fallout Radionuclide Measurements

The brief overview of field and laboratory techniques provided above has emphasized the potential uncertainty introduced into the application of fallout radionuclide measurements by sampling variability and measurement precision. Any measurement of the

fallout radionuclide activity or inventory is likely to involve a significant degree of uncertainty by virtue of both the heterogeneity of natural soils and sediments and the associated sampling variability and the limited precision of gamma spectrometry measurements. It is important that this uncertainty should be recognized, at least implicitly, if not explicitly, and if possible propagated through subsequent use of the data, for example to derive estimates of soil redistribution rates by comparing measurements of the fallout radionuclide inventory made at individual sampling points with the local reference inventory.

Some Examples of the Application of Fallout Radionuclides in Erosion and Sedimentation Investigations

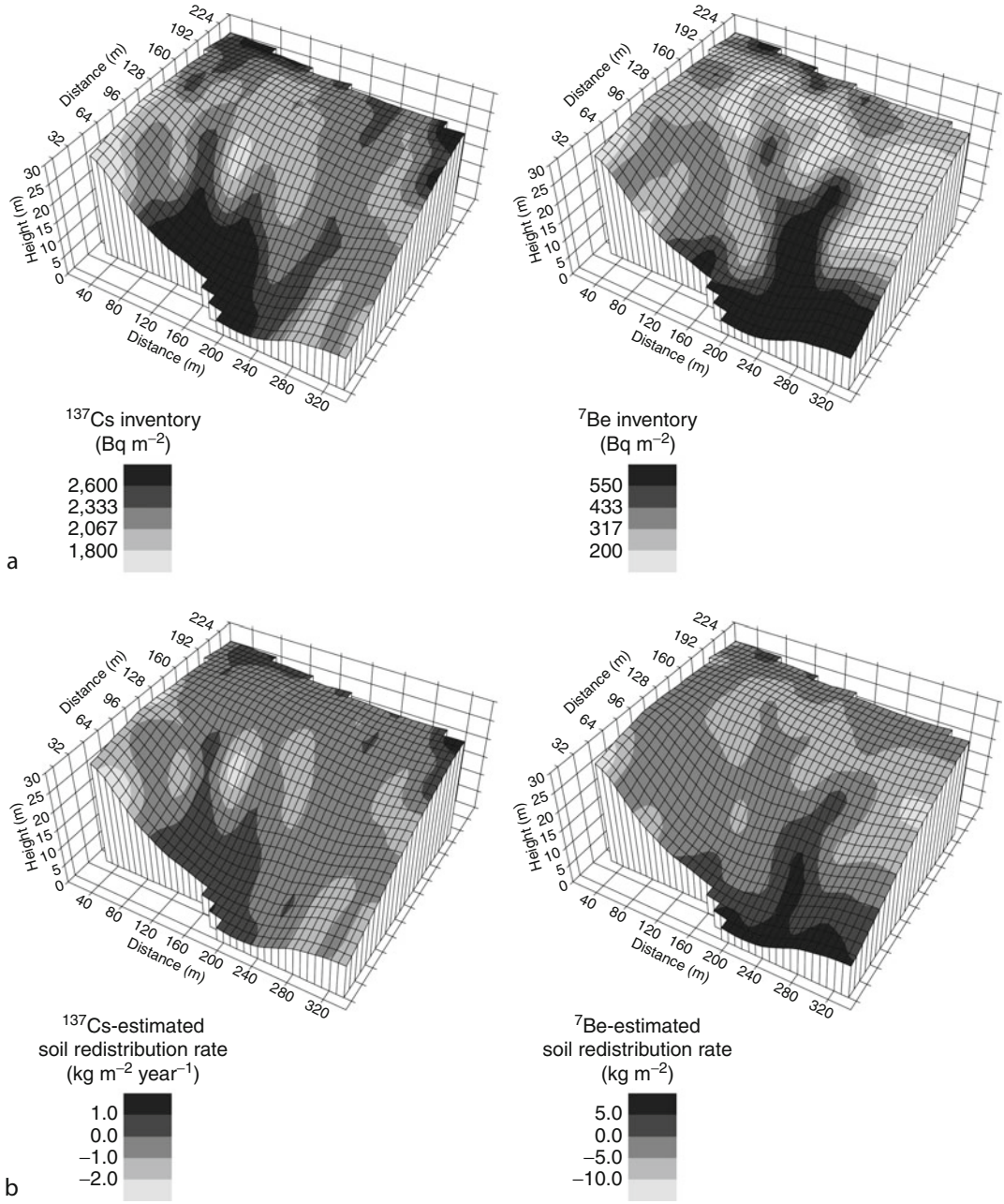
Further consideration of the application of fallout radionuclides in the study of erosion and sedimentation within the environment can usefully involve a brief overview of several case studies, based on investigations undertaken by the author and his coworkers, which demonstrate the potential of the approach across a range of applications. These include soil erosion, floodplain sedimentation, and sediment source tracing.

Soil Erosion: A Field Scale Investigation

Although most work involving the use of fallout radionuclides in soil erosion investigations has been based on measurements of ^{137}Cs , both $^{210}\text{Pb}_{\text{ex}}$ and ^7Be have also been used in similar applications. By collecting cores from a study site, measuring the ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$, or ^7Be inventories, and applying a conversion model, it is possible to derive point estimates of the erosion and deposition rates associated with the cores. The time base of the erosion rate estimates will depend on the radionuclide used. These values can then be used to establish the spatial pattern of soil redistribution, and by integrating the values across the study site, it is possible to calculate the gross and net erosion and the sediment delivery ratio. The latter represents the proportion of the eroded soil that is transported out of the study site and provides important information on the likelihood of eroded soil reaching watercourses. By making measurements of more than one radionuclide, it is often possible to obtain additional information, by

virtue of the different half-lives or fallout histories of the radionuclides. Figure 14 presents the results of an investigation of rates and patterns of soil redistribution within a 6.7-ha field at Higher Walton Farm near Crediton in Devon, UK [33]. The area covered by the field represents a valley head depression with slopes of up to 15° converging toward its outlet. In this case, measurements of both ^{137}Cs and ^7Be activities were undertaken, with the former providing estimates of average rates of soil redistribution over the past ~ 45 years, and the latter estimates of the erosion rates associated with a particular period of heavy rainfall (69 mm in 7 days) that occurred in early January 1998.

In this study, the soil cores used for the ^{137}Cs and ^7Be measurements were collected during two separate campaigns, although they could have been collected together. In both cases, the cores were collected at the intersections of a 20 m $\times$ 20 m grid, resulting in a suite of ~ 140 cores from each sampling campaign. Cores were also collected from adjacent areas of undisturbed land, which were not expected to have experienced either erosion or deposition, in order to establish the reference inventory. For the ^{137}Cs measurements, the cores were collected in August 1996, using a motorized percussion corer equipped with a 6.9-cm internal diameter steel core tube, which was inserted into the soil to a depth of ~ 60 cm. The cores used for the ^7Be measurements were, in contrast, much shallower and were collected manually to depths of 3–5 cm using a 15-cm diameter plastic core tube, in January 1998. During the preceding spring/summer of 1997, the field had been cultivated and sown to maize and the crop was harvested in early November 1997, when the soil was compacted by the harvesting equipment. After harvesting, the field was left bare and uncultivated over the winter and the period of heavy rainfall in early January 1998 resulted in substantial surface runoff and soil erosion. The pattern of ^{137}Cs inventories documented for the study field is presented in Fig. 14a. The cores collected from the adjacent undisturbed sites provided an estimate of the local ^{137}Cs reference inventory at the time of sampling of $\sim 2,500 \text{ Bq m}^{-2}$ and the pattern of ^{137}Cs inventories shows clear evidence of both erosion (reduced inventories) and deposition (increased inventories). Use of a mass balance conversion model incorporating the effects of soil redistribution by tillage enabled estimates of mean annual soil



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 14
The spatial distribution of ¹³⁷Cs and ⁷Be inventories within a field at Higher Walton Farm, near Crediton, UK (a) and of the estimates of soil redistribution rates derived from these inventories (b)

redistribution rates to be derived from the measured inventories. These soil redistribution rates relate to the period of ca. ~45 years prior to the sampling campaign in 1996. The pattern shown has been mapped in

Fig. 14b and the data have been summarized in Table 3, which presents values for the range of the documented soil redistribution rates, the mean erosion rate for the eroding areas, the mean deposition rate for the

depositional areas, the net soil loss from the field, and the sediment delivery ratio.

The spatial distribution of ^7Be inventories within the study field measured at the end of the period of heavy rainfall in early January 1998 is also presented in Fig. 14a. The equivalent value for the local reference inventory was estimated to be 533 Bq m^{-2} and the pattern depicted provides clear evidence of areas with both reduced and increased inventories, which in turn reflect both erosion and deposition within the field. In order to interpret this pattern in terms of soil redistribution rates associated with the period of heavy rainfall in early January, it was important to consider the extent to which it might reflect spatial variability inherited from previous erosion events. However, the preceding autumn and early winter had been relatively dry and there was no evidence of surface erosion having occurred during the previous 6 months. In this situation, the ^7Be inventory within the field immediately prior to the period of heavy rainfall in early January 1998 could be expected to be spatially uniform, since it would reflect the wet fallout associated with rainfall over the preceding months. Any spatial variability produced by soil redistribution associated with erosion events much earlier in 1997 would have been rendered insignificant by radioactive decay, due to the short half-life of ^7Be . It was therefore possible to assume that the spatial variability in ^7Be inventories within the study field evident in Fig. 14a reflects soil redistribution associated with the period of heavy rainfall in early January 1998. By using the conversion model for ^7Be measurements described previously in this chapter (i.e., Eq. 7), which takes account of the increase or decrease in inventory relative to the reference inventory and the depth distribution of ^7Be in uneroded soil within the field, it was possible to estimate the soil redistribution rates associated with the period of heavy rainfall preceding the sampling campaign. The pattern of soil redistribution rates demonstrated by these results is presented in Fig. 14b and summary data, equivalent to that provided for the ^{137}Cs measurements, are listed in Table 3. The soil redistribution rates (kg m^{-2}) associated with the period of heavy rainfall in early January 1998 estimated from the ^7Be measurements are substantially higher than the equivalent longer-term mean annual soil redistribution rates estimated using the ^{137}Cs measurements. However, the sediment delivery

Fallout Radionuclides and the Study of Erosion and Sedimentation. Table 3 A comparison of rates of soil redistribution within a field at Higher Walton Farm, near Crediton, UK, estimated using the ^{137}Cs and ^7Be measurements made on soil cores collected from the field

Measure	^{137}Cs ($\text{kg m}^{-2} \text{ year}^{-1}$)	^7Be (kg m^{-2})
Range	−4.5 to 2.0	−11.9 to 9.8
Mean erosion rate for eroding area	−1.1	−5.3
Mean deposition rate for depositional area	0.69	4.0
Net soil loss	−0.48	−2.5
Sediment delivery ratio	0.83	0.80

Based on Walling et al. [33]

ratios are closely similar, indicating that $\sim 80\%$ of the eroded sediment was transported out of the field. The high sediment redistribution rates associated with the period of heavy rainfall in early January 1998 reflect both the extreme nature of this period of rainfall and, perhaps more importantly, the condition of the field, which having been compacted by the maize harvesting machinery and left bare after the harvest was particularly susceptible to surface runoff and erosion. Such results underscore the potential significance of a small number of extreme events and the incidence of particular land use conditions in controlling erosion from the study field.

Soil Erosion: A Countrywide Survey

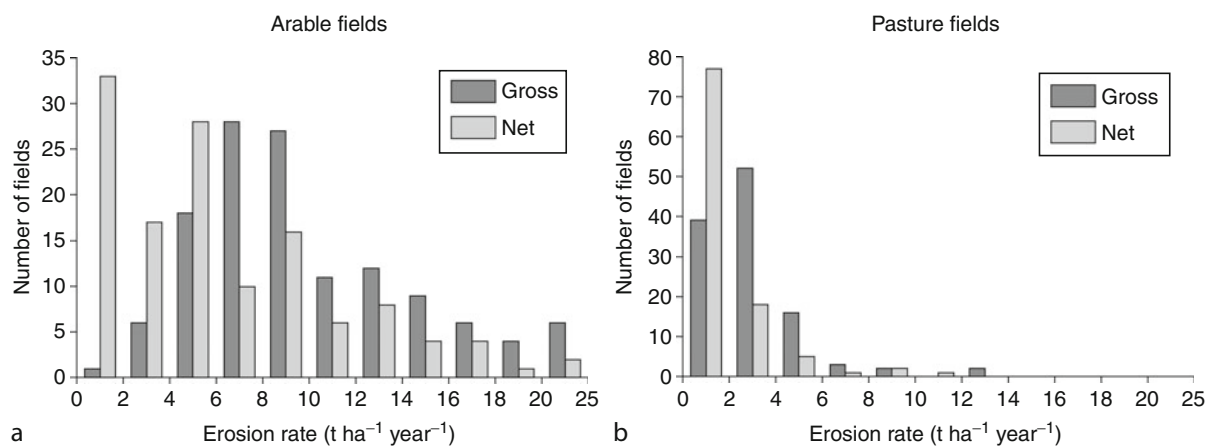
A key feature of most existing studies involving the use of measurements of ^{137}Cs and other fallout radionuclides in soil erosion investigations has been the emphasis on small areas, frequently individual fields. It has been argued that if fallout radionuclides are to be more widely used as an operational tool to support policy, as distinct from a research tool, there is a need to upscale their application. Analytical constraints, including the relative long count times and the substantial cost of gamma spectrometers, which inevitably

limit the number of detectors available in most laboratories, means that it is not feasible to upscale by simply collecting and analyzing a greatly increased number of samples. There is a need to devise novel sampling strategies that permit the use of ^{137}Cs measurements at the reconnaissance scale. Work by Loughran and coworkers [92] in Australia attempted to develop a national soil erosion survey based on ^{137}Cs measurements, and, in this case, 206 sites located in many different parts of the country were sampled using single downslope transects comprising about 20 sampling points. Recent work in the UK [158] has attempted to develop this approach further by reducing the number of samples collected from a site, while still obtaining a representative estimate of the gross and net erosion rate and thereby increasing the number of sites for which data can be assembled. By selecting fields of similar size and simple topography, but representative of a wide range of conditions known to influence soil erosion rates, including soil texture, slope steepness, effective rainfall and land use, it was possible to commence the assembly of a national scale database on soil erosion rates on agricultural land.

In the initial stages of this study, detailed investigations were undertaken in a number of fields to confirm the potential for using a small number of samples (e.g., [10–15]) collected along a carefully positioned downslope transect to provide a meaningful assessment of the gross and net erosion rates in a given field.

The information on ^{137}Cs inventories provided by the transects was used to estimate soil redistribution rates, using a conventional mass balance model and the diffusion and migration model for cultivated and non-cultivated fields, respectively. Estimates of the gross and net erosion rates for individual fields were derived by integrating the resulting estimates of soil redistribution rate along the length of the transect. Comparison of the results provided by the single transect with those provided by much denser grid or transect sampling networks, using the same conversion models, confirmed the validity of the approach and indicated that a transect comprising ca. 10–15 sampling points was capable of providing a reliable estimate of the gross and net erosion rates within individual fields.

A national sampling campaign was subsequently undertaken to obtain estimates of gross and net erosion rates for 248 fields located in different areas of England and Wales. The data collected are summarized in Fig. 15, which distinguishes the erosion rate estimates obtained for arable and pasture fields. These data have provided a basis for assessing the range of erosion rates occurring on agricultural land in England and Wales. The gross and net erosion rates documented for arable fields ranged between 1.2 and 29.3 $\text{t ha}^{-1} \text{ year}^{-1}$ and 0.0 and 27.3 $\text{t ha}^{-1} \text{ year}^{-1}$, with median rates of 6.6 and 5.2 $\text{t ha}^{-1} \text{ year}^{-1}$, respectively. The equivalent ranges for pasture were 0.7–13.0 and 0.0–11.7 $\text{t ha}^{-1} \text{ year}^{-1}$, with median values of gross and net erosion of 2.6 and



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 15

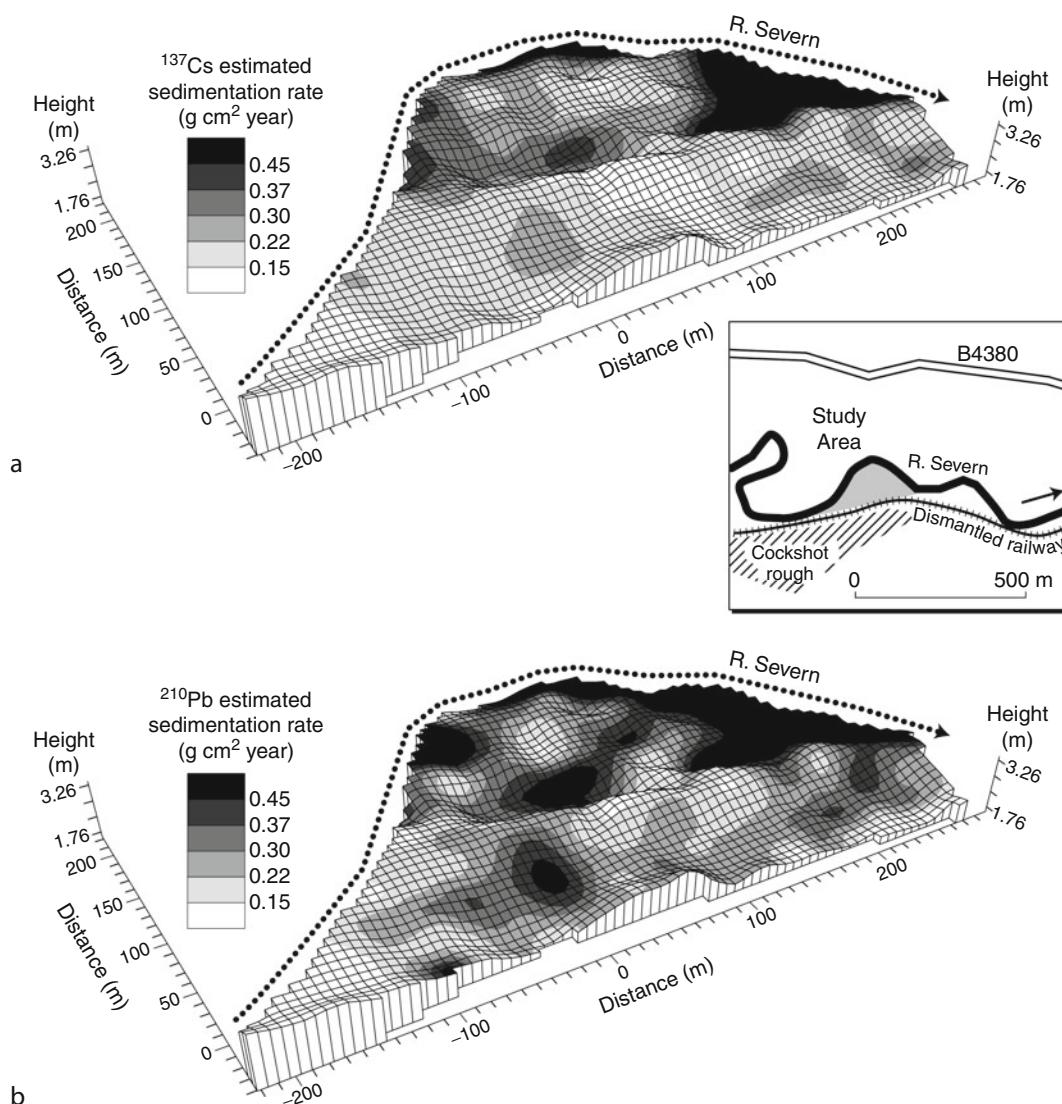
Frequency distributions of mean annual gross and net erosion rates for the arable and pasture fields in England and Wales sampled in the national survey undertaken by Walling and Zhang (2010) [158]

1.8 t ha⁻¹ year⁻¹, respectively. The ratio of net to gross erosion, which provides a useful measure of potential sediment delivery to watercourses, was typically >0.6 for arable fields and <0.6 for pasture fields, emphasizing the lower efficiency of sediment delivery from pasture areas. Comparison of the erosion rates obtained from the study with existing erosion hazard assessments, based on soil texture, confirmed the broad consistency of the results. The assembled dataset also provided a basis for exploring the influence of a range of potential controlling variables on soil erosion rates, using statistical tests. No very strong relationships existed, but several statistically significant relationships between the erosion rate variables and slope gradient and soil texture were found. The data have also been used to develop a typology for extrapolating the assembled information on gross and net erosion rates, based on land use, soil texture, and slope steepness and this has been employed within a GIS to produce national scale maps of gross and net rates of soil loss from agricultural land in England and Wales and a preliminary national scale inventory of soil loss for England and Wales.

Documenting Floodplain Sedimentation Rates

Overbank deposition on river floodplains during flood events represents an important potential sink for suspended sediment transported through a river system, and recent studies have demonstrated that such conveyance losses can be as high as ~40% of the suspended sediment load delivered to the main channel system [159]. Where attempts are made to reduce downstream sediment loads through improved land management and other sediment control measures, it can be important to document rates of overbank sedimentation on river floodplains, in order to quantify the potential magnitude of this sediment sink. In addition to representing a conveyance loss, the sediment deposited on the floodplain can also give rise to environmental problems, if it contains high levels of contaminants, such as heavy metals, or nutrients, particularly phosphorus. Again, information on rates of overbank sedimentation can represent an important requirement to assess such hazards. Since overbank inundation is commonly infrequent and difficult to forecast, direct measurements of associated sedimentation rates, using

sediment traps and related techniques, can prove difficult to implement. Furthermore, there will often be a need to establish longer-term rates of sedimentation to take account of the frequency of floods of different magnitude. In this context, the use of ¹³⁷Cs and ²¹⁰Pb_{ex} measurements to document longer-term mean annual sedimentation rates offers considerable potential. As an example, Fig. 16 shows how ¹³⁷Cs and ²¹⁰Pb_{ex} measurements were used to document overbank sedimentation rates along a short reach of the floodplain of the River Severn near Buildwas in Shropshire, UK. In this study, 124 sediment cores were collected at the intersections of a 25 m × 25 m grid using a motorized percussion corer equipped with a 6.9-cm internal diameter core tube. Cores were collected to a depth of ~70 cm to ensure that they included the complete ¹³⁷Cs and unsupported ²¹⁰Pb profiles. Measurements of the ¹³⁷Cs and ²¹⁰Pb_{ex} inventories of the cores were used to estimate the mean annual sedimentation rates at the coring points using the procedures documented by [118, 122] and described previously. In the case of ¹³⁷Cs, the estimate of the sedimentation rate derived from a sectioned core and the depth of the ¹³⁷Cs peak (see Eq. 8) was extrapolated to the other sampling points where bulk cores were collected, using the procedure represented in Eq. 10. In the case of the ²¹⁰Pb_{ex} measurements, the values of total inventory obtained for the individual bulk cores collected from the sampling points were used with the CICC model (Eq. 15) to estimate the sedimentation rates at those points. These estimates of sedimentation rate were used to map the patterns of sedimentation within the reach presented in Fig. 16. These patterns reflect the interaction of a number of controls on the sedimentation rate, including depth of inundation, distance from the channel, and the local microtopography. It is also possible to compare the sedimentation rate estimates derived from the ¹³⁷Cs measurements, which relate to the past ~40 years, with those based on the unsupported ²¹⁰Pb measurements, which relate to the past ~100 years, in order to assess longer-term changes in sedimentation at this location. The mean annual sedimentation rate at this site over the past 40 years is 0.28 g cm⁻² year⁻¹, whereas the equivalent rate for the past 100 years is 0.33 g cm⁻² year⁻¹. This suggests that rates of overbank sedimentation have decreased slightly toward the present. This decrease could reflect changes in the



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 16

The spatial distribution of overbank sedimentation rates within a small reach of the floodplain of the River Sever near Buildwas, Shropshire, UK derived from ^{137}Cs and $^{210}\text{Pb}_{\text{ex}}$ measurements undertaken on floodplain cores

magnitude and frequency of flood inundation, but it might also reflect a progressive reduction in inundation depths as the level of the floodplain surface gradually increased in height as a result of overbank deposition.

Using Floodplain Deposits to Reconstruct Changes in the Properties of Sediment Transported by a River

Where post-depositional changes in sediment geochemistry can be assumed to be of limited importance,

it is possible to use downcore changes in the geochemistry of overbank deposits to reconstruct changes in the properties of the suspended sediment transported by a river. The overbank deposits effectively preserve a record of the suspended sediment transported by the river in the recent past [160] and fallout radionuclides can be used to provide information on the chronology of such deposits. A useful example of this approach is provided by a study undertaken by the author and his coworkers on the floodplain bordering

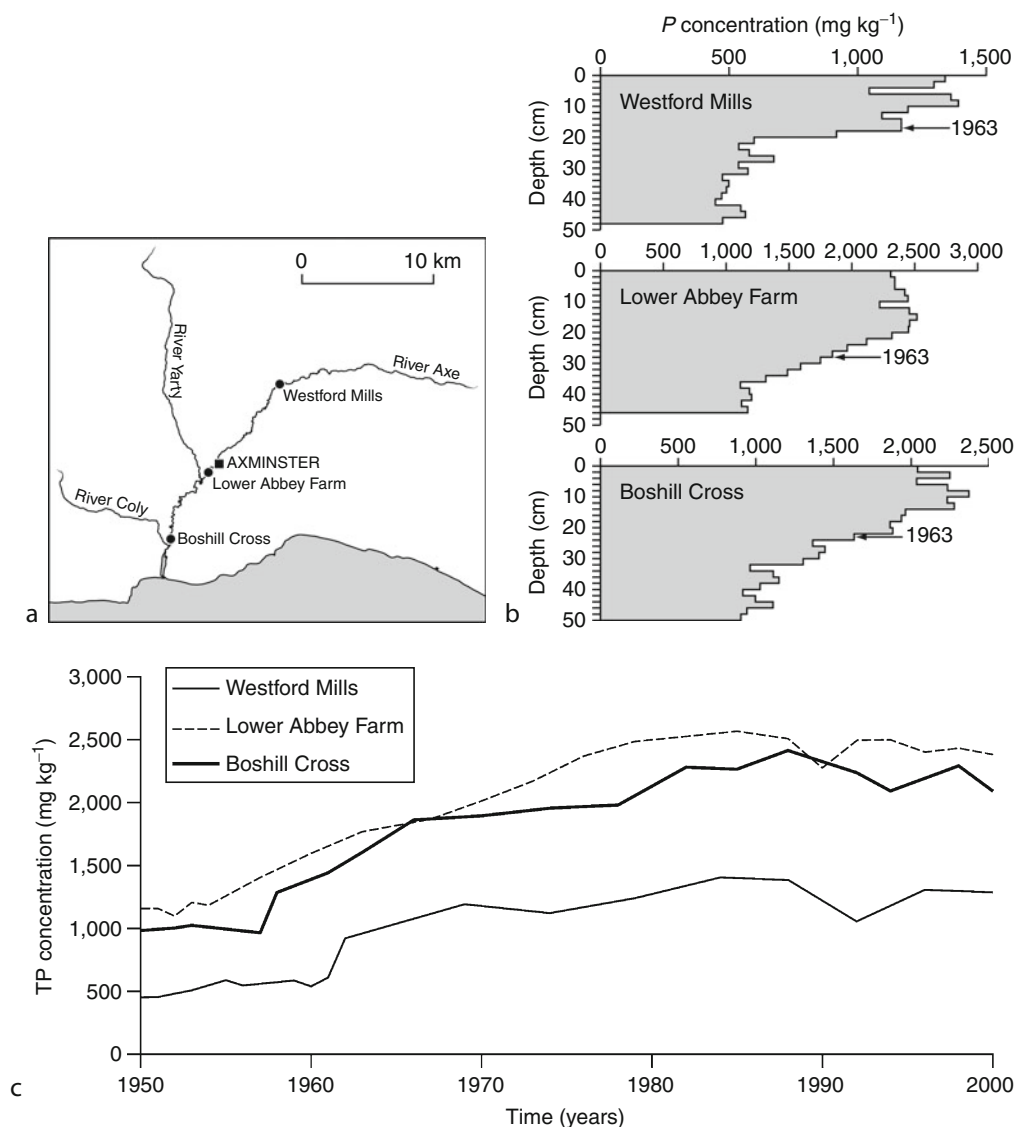
the main channel of the River Axe, which drains a 303 km² catchment in Devon, UK, aimed at reconstructing changes in the phosphorus content of the deposited sediment. Floodplain cores were collected from three representative sites in the upper, middle, and lower reaches of the river basin (see Fig. 17a) and measurements of both the ¹³⁷Cs activity and the total-P content of the sectioned cores were undertaken, in order to establish the associated depth profiles. The total-P profiles shown in Fig. 17 indicate that the total-P content of sediment deposited on the river floodplains at the three sites has more than doubled in recent decades, in response to increased fertilizer application and livestock densities in the upstream catchments and increased effluent discharges from sewage works and related point sources. The ¹³⁷Cs measurements were used to establish an approximate chronology for these changes by defining the level of the floodplain surface in 1963, the year of peak ¹³⁷Cs fallout (Fig. 17b), and assuming a constant annual sedimentation rate. This chronology was used to derive the plots of the changing total-P content of deposited sediment since 1950 presented in Fig. 17c. At all three sites, the total-P content of deposited sediment has increased by more than 100% over the past 50 years, but higher concentrations are found in the middle and lower reaches of the catchment where agriculture is more intensive and effluent discharges are greater. Although not clearly defined, there is some evidence of a reduced total-P content in deposited sediment in the last 10 years and this could reflect a response to improved land management and treatment of waste water discharges. Information of this nature can afford a valuable basis for defining baseline levels for fluvial sediment properties and assessing the impact of changing land management and wastewater treatment in reducing the inputs of P and other sediment-associated nutrients and contaminants to river systems.

Use of Fallout Radionuclides for Sediment Source Fingerprinting: An Example from the UK

The siltation of spawning gravels has been increasingly identified as a key factor contributing to the declining success of salmon fisheries in British rivers. Concern for this problem has focused attention on the need to reduce gravel siltation and thus to reduce fine sediment

mobilization and transport in impacted catchments, through the establishment of sediment control programs. The development of effective sediment control programs requires information on the likely sources of the fine sediment accumulating in spawning gravels, since these sources must be targeted if the control program is to prove effective. In an attempt to provide such information, a reconnaissance source fingerprinting survey of several representative rivers, located in different parts of Britain, was undertaken by the author and his coworkers in collaboration with the Environment Agency [161].

Samples of interstitial fine sediment were recovered from salmonid spawning gravels within a representative selection of rivers in England and Wales (Fig. 18), during a national fieldwork program conducted by the Environment Agency over the period 1999–2000. Sample collection was based on the use of retrievable basket samplers, which were installed in artificial redds constructed in spawning gravels at representative locations. Between one and five basket samplers were installed in each of the rivers identified on Fig. 18. The basket samplers were filled with clean framework gravel (>6.4 mm) prior to their installation, and they were retrieved ca. 3 months later. The gravel contained within the basket was subsequently wet sieved to recover the fine (<0.125 mm) interstitial sediment that had accumulated within the gravel during the period of deployment and this fraction was used for sediment source fingerprinting. By virtue of the reconnaissance nature of the study, which included 18 catchments, sampling of potential source materials focused on the broad distinction between surface and channel bank/subsurface sources, and a total of 672 source material samples were collected from the different study areas. These samples were sieved to <0.125 mm to facilitate direct comparison with the samples of fine interstitial sediment. The limited resources available to the study also precluded the use of an extensive range of fingerprint properties to discriminate the two potential sources and emphasis was placed on the use of radiometric (¹³⁷Cs, ²¹⁰Pb_{ex}, ²²⁶Ra) measurements, coupled with information on the organic properties (C and N content) of the potential sources. Mean values of these properties were used to characterize the two potential sources in each of the river basins investigated.

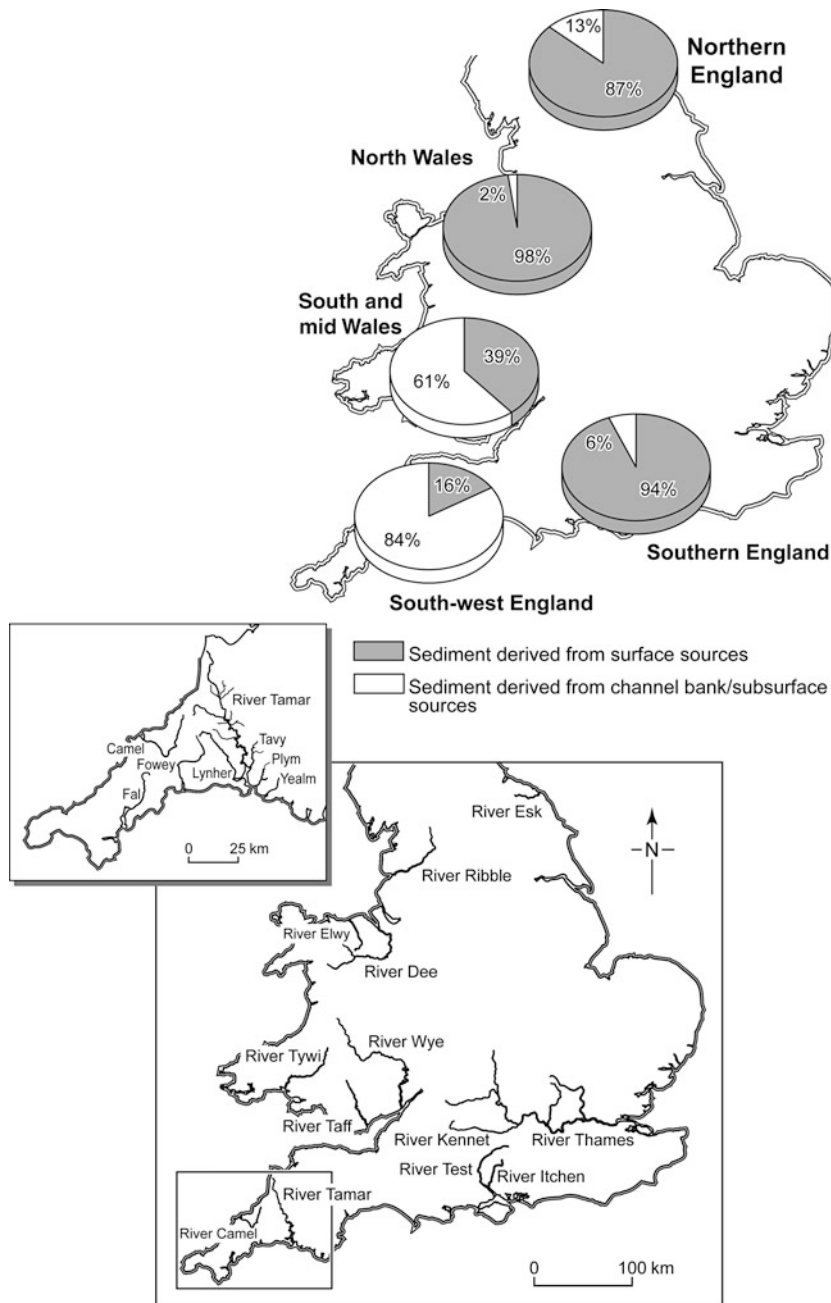


Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 17

The changing total-P content of fine sediment deposited on the floodplain of the River Axe over the past 50 years, showing (a) the location of the coring points, (b) the total-P depth distributions, and (c) the reconstructed trends in the total-P content of fine sediment deposited on the floodplain over the past 50 years

A multicomponent mixing model, incorporating correction for the effects of contrasts in particle size and organic matter content between the samples of interstitial sediment and the source material, was used to estimate the relative contributions of surface and channel bank/subsurface sources to the fine sediment recovered from the spawning gravels. The results of these computations for different regions of England

and Wales are presented in Fig. 18. Significant contrasts in the relative contributions of the two sources are apparent between the regions, with, for example, surface sources accounting for >90% of the fine interstitial sediment in north Wales and southern England, but only 16% and 39% in south-west England and south and mid Wales, respectively. These regional contrasts reflect the interaction of land use and both erosion



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 18

Fingerprinting the source of fine sediment accumulating in salmon spawning gravels, showing the location of the sampled rivers and regional contrasts in the source of the fine sediment (Based on [161])

processes and the efficiency of sediment transfer to the channel network. In south-west England, where stocking densities are frequently high, drainage densities are relatively high and river channels are generally incised,

a combination of livestock trampling of channel banks and erosion of unstable well-developed vertical channel banks mean that channel and subsurface sources are the dominant source of fine interstitial sediment.

In contrast, the greater importance of arable farming and more specifically soil erosion on large cultivated fields with few boundaries to interrupt slope-channel connectivity, combined with the relative stability of the well-vegetated channel banks, result in surface soils providing the dominant source (94%) of the samples of fine interstitial sediment recovered from spawning gravels in southern England. Surface sources are also important in the upland areas of northern England and north Wales, where high rainfall and grazing pressure promote the erosion of surface soils under upland pasture or moorland and the steep topography combined with an absence of field boundaries results in the efficient routing of sediment to the river channels.

The contrasts in the relative importance of surface and channel bank/subsurface sources in different regions of England and Wales outlined above have important implications for the design and implementation of effective sediment control strategies. Where bank erosion is the dominant source, attention should clearly focus on reduction of livestock trampling of channel margins and improvement of bank stability (e.g., by fencing and revegetation of channel margins). However, such measures are likely to be of little value in areas where surface sources are dominant and where emphasis should be placed on controlling sediment mobilization and transfer within the catchment more generally.

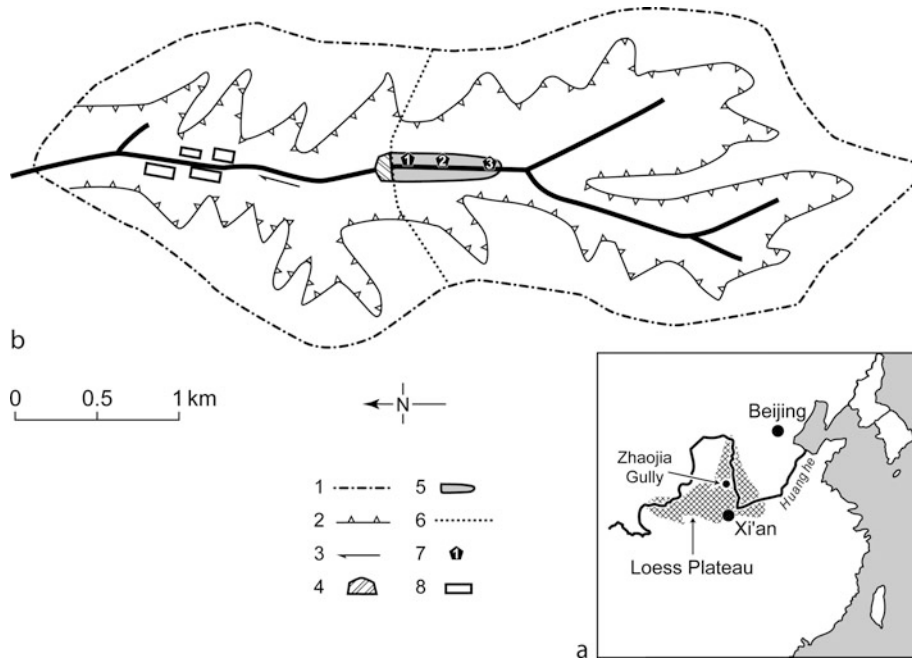
Use of Fallout Radionuclides for Sediment Source Fingerprinting: An Example from China

Another example of the potential for using environmental radionuclides, and more particularly ^{137}Cs , for fingerprinting sediment sources is drawn from work in the rolling loess plateau region of Shaanxi Province, China [162]. The landscape in the study area comprises two contrasting geomorphological units, representing the rolling plateau surface and the gully systems which dissect the plateau surface. The very high sediment yields in this environment ($\sim 10,000 \text{ t km}^{-2} \text{ year}^{-1}$) cause major downstream problems in terms of reservoir and channel siltation, and this has focused attention on the need to reduce sediment yields through soil and water conservation and sediment control measures. In planning such measures, there is a need to

establish the relative importance of soil erosion from the cultivated areas of the rolling plateau and erosion occurring within the gully systems, in contributing to downstream sediment yields, since these two potential sediment sources require very different approaches to reduction in sediment mobilization and transfer.

A study aimed at establishing the relative importance of the two main sediment sources was undertaken in the 3.86 km^2 Zhaojia gully catchment, where the plateau surface and gully systems were of similar spatial extent and accounted for 53% and 47% of the catchment area respectively (Fig. 19). A check dam had been constructed across the channel in the middle portion of the catchment in early 1991 (Fig. 19), and when the catchment was visited in October 1993, the deposits associated with two floods occurring during the period July–August 1993 were clearly distinguishable in sediment cores collected from the area of sediment deposition above the check dam. The total depth of sediment deposition associated with the two events was $\sim 33 \text{ cm}$. Based on an analysis of three cores, the mean ^{137}Cs content of sediment deposited by the first flood was 0.74 Bq kg^{-1} , whereas that for the second flood was 1.06 Bq kg^{-1} . In attempting to decipher the main sources of the deposited sediment and thus the sediment yield from the catchment, attention was directed to variations in the ^{137}Cs concentration of potential sources within the two main sediment sources areas. Mean ^{137}Cs concentrations associated with these sources exhibited significant variations and offered an effective means of discriminating the sources. A summary of the ^{137}Cs content of potential source materials in the Zhaojia Gully catchment is provided in Table 4.

When the average ^{137}Cs content of the sediment deposits associated with the two floods occurring in 1993 was compared with that of potential source material collected shortly after the flood (cf. Table 4), it was found to be similar to that of cultivated soil from the steep gully slopes, lower than that of soil from the cultivated land on the plateau crests and steeper lower slopes of the rolling plateau, and considerably lower than that of topsoil from the grass covered areas. It was also very substantially higher than that of surface material from steep bare slopes and from recent mass movement deposits within the gully area. In view of either



Fallout Radionuclides and the Study of Erosion and Sedimentation. Figure 19

The Zhaojia Gully catchment showing its location (a) and its main features (b). 1, catchment divide; 2, the boundary between the rolling plateau and the gully areas; 3, the main channel; 4, the check dam; 5, area of sediment deposition associated with the check dam; 6, divide of the catchment contributing to the sediment trap reservoir; 7, sampling sites for sediment deposits; 8, village (Based on [162])

Fallout Radionuclides and the Study of Erosion and Sedimentation. Table 4 The ^{137}Cs content of potential source materials in the Zhaojia Gully catchment from different topographic locations and under different land use conditions

Zone	Source material and land type	Catchment area (in %)	No. of samples	Mean ^{137}Cs content (Bq kg^{-1})	Erosional behavior
Gully (47%)	Surface soil on grassland of gully slopes	21	8	7.0	Slight sheet and rill erosion
	Plowed soil on steep gully slopes	2	7	1.2	Very severe sheet and rill erosion
	Surface soil on bare slopes and mass movement deposits	19	7	0.02	Very severe sheet and rill erosion on bare slopes and active mass movements
	Village area, cultivated land on terrace and gully floor and tracks	5	–	–	
Rolling plateau (53%)	Cultivated soil of crest slopes	14	36	6.4	Moderately severe sheet and rill erosion
	Cultivated soil of lower slopes	36	32	3.4	Severe sheet and rill erosion
	Surface soil on grassland	3	–	–	

Based on Walling et al. [162]

their limited extent or their low erosion rates, the cultivated land on the steep gully slopes and the grassland within the gully area were deemed unlikely to represent important sediment sources and the cultivated areas of the rolling plateau and the steep bare slopes and recent mass movement deposits within the gully area were identified as representing the primary sediment sources. Taking account of the relative extents of the cultivated areas on the crests and the steeper upper slopes of the rolling plateau and estimates of erosion rates obtained from these areas using ^{137}Cs measurements, a weighted average ^{137}Cs content for sediment mobilized from these cultivated areas of the rolling plateau of 3.9 Bq kg^{-1} was estimated.

Assuming that the average ^{137}Cs content of surface material from the steep bare gully slopes and recent mass movement deposits (0.02 Bq kg^{-1}) was representative of sediment from the gully area and that the minor contributions from the cultivated areas and areas of grass cover within the gully could be ignored, a simple mixing model was used to estimate the relative contribution of sediment derived from the gully area and the cultivated areas of the rolling plateau to the flood deposits. The results provided by the mixing model are presented in Table 5. These indicate that during the first flood, 18.6% of the sediment was derived from the rolling plateau and 81.4% from the gully area. During the second event, the equivalent values were 26.8% and 73.2%. Taking the two events together, the values were 23% from the rolling plateau and 77% from the gully area. These results emphasized the importance of the gully area and the associated gully erosion and mass movement processes as the main sediment source within the Zhaojia Gully catchment and clearly demonstrated that any strategy aimed at reducing sediment yields from this catchment and

from similar landscapes in the local area should focus on reducing sediment production within the gully area. Such measures could include check dams and other gully stabilization measures.

Future Directions

As demonstrated by the above overview, the use of fallout radionuclides in soil erosion and sedimentation investigations now has a history that stretches back over more than 30 years. Over this period, there have been advances in measurement technology and whereas early work primarily focused on the use of ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$ and ^7Be are now also quite widely used both as an alternative and a complement to ^{137}Cs . Equally, whereas most early work focused on documenting patterns and rates of soil erosion and establishing chronologies for lake cores, subsequent work has expanded the various applications to include other sediment sinks, particularly river floodplains and wetlands, and sediment source fingerprinting. These expanding applications have been promoted by two key drivers. The first is the growing recognition of both the many environmental problems associated with the mobilization and transfer of fine sediment within the landscape and the need to document and develop an improved understanding of the source, mobilization, and transfer of fine sediment, in order to develop effective management and control strategies. The second is the essentially unique information that can be provided by the use of fallout radionuclides as tracers and the need to develop the use of tracers to complement more traditional monitoring techniques [163]. Future developments are likely to expand further this field of work and four areas in particular can be identified as potential areas for expansion or advances.

Fallout Radionuclides and the Study of Erosion and Sedimentation. Table 5 The relative contribution of sediment from the rolling plateau and gully areas of the Zhaojia Gully catchment during the floods of 1993

Flood	Date of flood deposit	Deposit thickness (cm)	^{137}Cs content of sediment deposits (Bq kg^{-1})	Relative contribution (%)	
				Rolling plateau	Gully
2	August 21, 1993	18.3	1.06	26.8	73.2
1	July 21–August 1, 1993	14.7	0.74	18.6	81.4

Based on Walling et al. [162]

These include: firstly, the use of other radionuclides; secondly, the further development and improvement of measurement techniques; thirdly, greater integration of the various applications and approaches; and finally, the progressive incorporation of fallout radionuclide tracers into sediment-based catchment investigations, as an operational rather than as a research tool. Each of these future directions will be briefly considered.

Use of Other Radionuclides

^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$ and ^7Be are now firmly established as the key fallout radionuclides for use in sediment tracing investigations. Their different half-lives and contrasting fallout patterns usefully complement each other. However, in recent years, some attention has been given to the potential for using Plutonium-239 and 240 ($^{239,240}\text{Pu}$) as an alternative to ^{137}Cs [164–167]. These radionuclides are also a product of bomb testing and have been shown to demonstrate very similar behavior to ^{137}Cs in terms of their distribution within the soil profile [168], their post-fallout redistribution within the landscape and thus their potential for studying soil redistribution, [169, 170], dating sediment cores [167, 171, 172], and fingerprinting sediment sources [166]. However, some contrasts with ^{137}Cs in terms of their post-fallout redistribution have been reported [173] and it has been suggested that $^{239,240}\text{Pu}$ is likely to be less mobile than ^{137}Cs in sediment cores [167].

The likely advantages of $^{239,240}\text{Pu}$ over ^{137}Cs lie primarily in their longer half-lives (24,110 and 6,561 years, respectively) and secondarily in analytical requirements. In the first context, the relatively short half-life of ^{137}Cs (30.1 years) means that activities of this radionuclide are beginning to decline significantly in the environment, since it is now almost 50 years since the period of peak fallout around 1963. Where fallout fluxes were relatively low, as, for example, in the southern hemisphere, this can represent a problem for gamma spectrometry analysis. Reduced activities mean longer count times and are likely to be coupled with reduced measurement precision and increased minimum levels of detection. With their long half-lives, $^{239,240}\text{Pu}$ do not face this problem. In addition, $^{239,240}\text{Pu}$ can provide additional chronomarkers associated with changes in the Pu isotopic composition of

fallout during the 1950s and 1960s. In particular, the $^{238}\text{Pu}/^{239+240}\text{Pu}$ activity ratio shows distinct shifts in the early 1950s and the mid-to-late 1960s, providing important known-age horizons in southern hemisphere sediments. In the second context, recent advances in analytical techniques have provided the possibility of analyzing $^{239,240}\text{Pu}$ by mass spectrometry (ICPMS and AMS), thereby providing a means of greatly increasing sample throughput, relative to ^{137}Cs , where individual samples collected in areas with low fallout inventories can require a count time of ~ 24 h or more. A further feature of $^{239,240}\text{Pu}$ fallout is that, unlike ^{137}Cs , it was not associated with both bomb and Chernobyl fallout. It was only associated with the former. In some areas of the world affected by Chernobyl fallout, $^{239,240}\text{Pu}$ may therefore offer opportunities to distinguish fallout inputs from both sources [174] and to exploit their differences in timing to obtain additional information on the erosional history of a study area.

Another radionuclide which has attracted attention for its use as an alternative to ^{137}Cs in dating lake cores, due to the likelihood of reduced mobility, is ^{241}Am . ^{241}Am is a product of the decay of ^{241}Pu from bomb fallout and therefore has the advantage that activities are currently increasing rather than decreasing.

Development and Improvement of Measurement Technology

It is possible that advances in technology might also lead to alternatives to gamma spectrometry for the analysis of ^{137}Cs , ^7Be , and ^{210}Pb . The potential to increase sample throughput would offer considerable scope for increasing the size of the area that can be investigated and/or the spatial resolution of the results. Other areas of measurement technology, where advances have begun to occur, include the remote sensing of spatial patterns of radionuclide inventories using airborne sensors. The successful use of this approach for ^{137}Cs has been reported in Northern England and Northern Ireland [175, 176]. As yet, the spatial resolution is limited and the approach is more suited to documenting spatial variability of the original fallout input rather than post-fallout redistribution. Furthermore, its use may be restricted to areas receiving additional Chernobyl fallout where ^{137}Cs inventories

are higher. However, the possibility of collecting detailed information on the spatial distribution of inventories across an area would clearly introduce important new opportunities for expanding the scale of sediment redistribution investigations.

Increased Integration of Radionuclides and Approaches

Two important features of the use of fallout radionuclides as tracers in erosion and sedimentation investigations are: firstly, that conjunctive use of different radionuclides can provide increased information on the functioning of the system under investigation, and secondly, that it is possible to use the same tracers to document various components of a catchment sediment budget, for example, soil redistribution rates on the slopes; the deposition and storage of fine sediment on downstream floodplains; the dynamics of sediment mobilization, deposition, and remobilization within the channel system; and the relative contribution of key sediment sources to the sediment output from the catchment. Furthermore, where a long-term sink, such as a lake or well-developed floodplain or wetland, exists at the outlet of the catchment, it is possible to use the same radionuclides to establish a chronology for sediment cores to assist in interpreting the erosional history of the catchment. The potential for conjunctive use of several radionuclides is being increasingly recognized, and Blake and coworkers [177] provide an interesting example of this potential in their study of sediment redistribution by erosion after a wildfire at a site in Australia. Equally, the potential for integrating work on several different components of the sediment budget is provided by Zhang and Walling [178] who showed how it is possible to use the information provided by ^{137}Cs depth distributions in lake and reservoir sediments to provide information on both erosion rates in the contributing catchments and the dominant sediment sources.

To date, however, studies involving different components of the sediment budget have often been compartmentalized and considerable scope exists to integrate them more closely to provide a comprehensive assessment of the catchment sediment budget. In this context, there will be a need to make use of other measurements and approaches to complete the picture,

but fallout radionuclide measurements can provide the key to establishing the sediment budget of a catchment [163, 179–181]. As indicated previously, the catchment sediment budget represents a key source of information for designing and implementing effective sediment management strategies for catchments and river basins [182]. Further work is required to integrate the various applications of fallout radionuclides in sediment budget investigations and to develop the comprehensive toolbox that will be required to exploit their full potential.

Moving from a Research Tool to an Operational and Management Tool

To date, most of the studies aimed at exploring and developing the potential for using fallout radionuclides in studies of erosion and sedimentation have been curiosity-driven research investigations. In some instances, the starting point has been the potential application, which has then been demonstrated. In others, the starting point has been the need to solve a problem or to obtain an improved understanding of a particular part of the system, and fallout radionuclides have been successfully used as the tool to achieve that objective. However, as the comprehensive toolbox referred to above is progressively developed, it is to be hoped that the wide-ranging potential of the approach will gain increasing recognition and that the use of fallout radionuclides will become more routine and thus a more widely used tool to support the sustainable management of the environment.

Bibliography

Primary Literature

1. Oldeman LR (1994) The global extent of land degradation. In: Greenland DJ, Szabolcs I (eds) *Land resilience and sustainable land use*. CAB International, Wallingford, pp 99–118
2. Speth JG (1994) Desertification convention essential for food security. *Environ Conserv* 21:4–5
3. Toy TJ, Foster GR, Renard KG (2002) *Soil erosion: processes, prediction, measurement and control*. Wiley, New York, 352 p
4. Buringh P (1981) An assessment of the losses and degradation of productive agricultural land in the world. *FAO Working Group on Soils Policy*, Rome
5. Brown LR (1984) Conserving soils. In: Brown LR (ed) *State of the world 1984*. Norton, New York, pp 53–75

6. Yang D, Kanae S, Oki T, Koike T, Muslake K (2003) Global potential soil erosion with reference to land use and climate changes. *Hydrol Process* 17:2913–2928
7. Lal R (2003) Soil erosion and the global carbon budget. *Environ Int* 29:437–450
8. Van Oost K, Quine TA, Govers G, De Gryze S, Six J, Harden JW, Ritchie JC, McCarty GW, Heckrath G, Kosmas C, Giraldez JV, Marques da Silva JR, Merckx R (2007) The impact of agricultural soil erosion on the global carbon budget. *Science* 318:626–629
9. Clark EH, Haverkamp JA, Chapman W (1985) *Eroding soils: the off-farm impacts*. The Conservation Foundation, Washington, DC
10. Pimentel D, Harvey C, Resosuarmono P, Sinclair K, Kurz D, McNair M, Crist S, Shpritz L, Fitton L, Saffourni R, Blair R (1995) Environmental and economic costs of soil erosion and conservation benefits. *Science* 267:1117–1123
11. Mahmood K (1987) Reservoir sedimentation: Impact, extent and mitigation. World Bank Technical Paper no. 71. The World Bank, Washington, DC
12. Lal R (1994) *Soil erosion research methods*. St. Lucie Press, Boca Raton, 340 p
13. Morris GL, Fan J (1998) *Reservoir sedimentation handbook: design, and management of dams, reservoirs and watersheds for sustainable use*. McGraw Hill, New York, 750 p
14. Mabit L, Benmansour M, Walling DE (2008) Comparative advantages and limitations of the fallout radionuclides ^{137}Cs , $^{210}\text{Pb}_{\text{ex}}$ and ^7Be for assessing soil erosion and sedimentation. *J Environ Radioactiv* 99:1799–1807
15. Goldberg ED (1963) Geochronology with Pb-210 in radioactive dating. In: *Proceedings of a symposium on radioactive dating*, International Atomic Energy Agency, Vienna. Contribution 1510, pp 121–131
16. Menzel RG (1960) Transport of strontium-90 in runoff. *Science* 131:499–500
17. Yamagata N, Matsuda S, Kodaira K (1963) Run-off of caesium-137 and strontium-90 from rivers. *Nature* 200:668–669
18. Rogowski AS, Tamura T (1965) Movement of ^{137}Cs by runoff, erosion and infiltration on the alluvial Captina silt loam. *Health Phys* 11:1333–1340
19. Ritchie JC, Spraberry JA, McHenry JR (1974) Estimating soil erosion from the redistribution of fallout ^{137}Cs . *Soil Sci Soc Am J* 38:137–139
20. McHenry JR, Ritchie JC (1977) Estimating field erosion losses from fallout Cs-137 measurements. International Association of Hydrological Sciences Publication no. 122. IAHS, Wallingford, pp 26–33
21. de Jong E, Begg CBM, Kachanoski RG (1983) Estimates of soil erosion and deposition from some Saskatchewan soils. *Can J Soil Sci* 63:607–617
22. Kachanoski RG, de Jong E (1984) Predicting the temporal relationship between cesium-137 and erosion rate. *J Environ Qual* 13:301–304
23. McCallan ME, O'Leary BM, Rose CW (1980) Redistribution of caesium-137 by erosion and deposition on an Australian soil. *Aust J Soil Sci* 18:119–128
24. Campbell BL, Loughran RJ, Elliott GL (1982) Caesium-137 as an indicator of geomorphic processes in a drainage system. *Aust Geogr Stud* 20:49–64
25. Loughran RJ, Elliott GL, Campbell BL, Shelly DJ (1988) Estimation of soil erosion from caesium-137 measurements in a small cultivated catchment in Australia. *Appl Radiat Isot* 39:1153–1157
26. Walling DE, Bradley SB (1988) The use of caesium-137 measurements to investigate sediment delivery from cultivated areas in Devon, UK. International Association of Hydrological Sciences Publication no. 174. IAHS, Wallingford, pp 325–335
27. Walling DE, Quine TA (1991) The use of ^{137}Cs measurements to investigate soil erosion on arable fields in the U.K.: potential applications and limitations. *J Soil Sci* 42:147–165
28. Golosov VN, Panin AV, Markelov MV (1999) Chernobyl Cs-137 redistribution in the small basin of the Lokna river, Central Russia. *Phys Chem Earth A* 24:881–885
29. Zapata F (ed) (2002) *Handbook for the assessment of soil erosion and sedimentation using environmental radionuclides*. Kluwer, Dordrecht, 219 p
30. Ritchie JC, Ritchie CA (2008) *Bibliography of publications of ^{137}Cs studies related to erosion and sediment deposition*. USDA-ARS Hydrology and Remote Sensing Laboratory Occasional Paper HRSL-2008-02
31. Wallbrink PJ, Murray AS (1996) Distribution of ^7Be in soils under different surface cover conditions and its potential for describing soil redistribution processes. *Water Resour Res* 32:467–476
32. Blake WH, Walling DE, He Q (1999) Fallout beryllium-7 as a tracer in soil erosion investigations. *Appl Radiat Isot* 51:599–605
33. Walling DE, He Q, Blake W (1999) Use of Be-7 and Cs-137 measurements to document short- and medium-term rates of water-induced soil erosion on agricultural land. *Water Resour Res* 35:3865–3874
34. Matisoff G, Bonniwell EC, Whiting PJ (2002) Soil erosion and sediment sources in an Ohio watershed using Beryllium-7, Cesium-137, and Lead-210. *J Environ Qual* 31:54–61
35. Wilson CG, Matisoff G, Whiting PJ (2003) Short-term erosion rates from a Be-7 inventory balance. *Earth Surf Process Land* 28:967–977
36. Wallbrink PJ, Murray AS (1996) Measuring soil loss using the inventory ratio of excess lead-210 to cesium-137. *Soil Sci Soc Am J* 60:1201–1208
37. Walling DE, He Q, Quine TA (1995) Use of caesium-137 and lead-210 as tracers in soil erosion investigations. International Association of Hydrological Sciences Publication no. 229. IAHS, Wallingford, pp 163–172
38. Walling DE, He Q (1999) Using fallout lead-210 measurements to estimate soil erosion on cultivated land. *Soil Sci Soc Am J* 63:1404–1412
39. Walling DE, Collins AL, Sickingabula HM (2003) Using unsupported lead-210 measurements to investigate soil erosion and sediment delivery in a small Zambian catchment. *Geomorphology* 52:193–213

40. Ritchie JC, McHenry JR, Gill AC, Hawks PH (1972) Fallout Cs-137 in reservoir sediments. *Health Phys* 22:97–98
41. McIntyre SC, Naney JW (1991) Sediment deposition in a forested inland wetland with a steep-farmed watershed. *J Soil Water Conserv* 46:64–66
42. Craft CB, Casey WP (2000) Sediment and nutrient accumulation in floodplain and depressional freshwater wetlands of Georgia, USA. *Wetlands* 20:323–332
43. DeLaune RD, Patrick WH Jr, Buresh RJ (1978) Sedimentation rates determined by ^{137}Cs dating in a rapidly accreting salt marsh. *Nature* 275:532–533
44. Oenema O, DeLaune RD (1988) Accretion rates in salt marshes in Eastern Scheldt, South-west Netherlands. *Estuaries Coast Shelf Sci* 26:379–394
45. Walling DE, He Q (1998) The spatial variability of overbank sedimentation on river floodplains. *Geomorphology* 24:209–223
46. Terry JP, Garimella S, Kostaschuk RA (2002) Rates of floodplain accretion in a tropical island river system impacted by cyclones and large floods. *Geomorphology* 42:171–182
47. Krishnaswamy S, Lal D, Martin JM, Meybeck M (1971) Geochronology of lake sediments. *Earth Planet Sci Lett* 11:407–414
48. Robbins JA, Edgington DN (1975) Determination of recent sedimentation rates in Lake Michigan with Pb-210 and Cs-137. *Geochim Cosmochim Acta* 39:285–304
49. Appleby PG, Oldfield F (1978) The calculation of lead-210 dates assuming a constant rate of supply of unsupported lead-210 to the sediment. *Catena* 5:1–8
50. Robbins JA, Edgington DN, Kemp LW (1978) Comparative ^{210}Pb , ^{137}Cs , and pollen geochronologies of sediments from Lake Ontario and Erie. *Quatern Res* 10:256–278
51. Carroll J, Lerche I (2003) *Sedimentary processes: quantification using radionuclides*. Elsevier, Oxford
52. Krishnaswami S, Benninger LK, Aller RC, Von Damm KL (1980) Atmospherically-derived radionuclides as tracers of sediment mixing and accumulation in near-shore marine and lake sediments: evidence from ^7Be , ^{210}Pb and $^{239,240}\text{Pu}$. *Earth Planet Sci Lett* 47:307–318
53. Blake WH, Walling DE, He Q (2002) Using cosmogenic beryllium-7 as a tracer in sediment budget investigations. *Geogr Ann A Phys Geogr* 84A:89–102
54. Walling DE, He Q (2000) The global distribution of bomb-derived ^{137}Cs reference inventories. Unpublished report to the International Atomic Energy Agency
55. Smith JT, Beresford NA (2005) *Chernobyl: catastrophe and consequences*. Praxis, Chichester, 305 p
56. Winkler R, Rosner G (2000) Seasonal and long-term variation of ^{210}Pb concentration in air, atmospheric deposition rate and total deposition velocity in South Germany. *Sci Total Environ* 263:57–68
57. Olsen CR, Larsen IL, Lowry PD, Cutshall NH, Todd JF, Wong GTF, Casey WH (1985) Atmospheric fluxes and marsh-soil inventories of ^7Be and ^{210}Pb . *J Geophys Res* 90:10487–10495
58. Valles I, Camacho A, Ortega X (2009) Natural and anthropogenic radionuclides in airborne particulate samples collected in Barcelona (Spain). *J Environ Radioactiv* 100:102–107
59. Preiss N, Mélières MA, Pourchet M (1996) A compilation of data on lead-210 concentration in surface air and fluxes at the air-surface and water-sediment interfaces. *J Geophys Res* 101:28847–28862
60. Beks JP, Eisma D, van der Plicht J (1998) A record of atmospheric ^{210}Pb deposition in the Netherlands. *Sci Total Environ* 222:35–44
61. Turekian KK, Nozaki Y, Benninger LK (1977) Geochemistry of atmospheric radon and radon producers. *Annu Rev Earth Planet Sci* 5:227–255
62. Yamamoto M, Sakaguchi A, Sasaki K, Hirose K, Igarashi Y, Kim CK (2006) ^{210}Pb and ^7Be deposition: features of the Japan Sea side of Japan. *J Environ Radioactiv* 55:121–123
63. Baskaran M, Coleman CH, Santschi PH (1993) Atmospheric depositional fluxes of Be-7 and Pb-210 at Galveston and College Station, Texas. *J Geophys Res* 98:20555–20571
64. Cailliet S, Arpagaus P, Monna F, Dominik J (2001) Factors controlling Be-7 and Pb-210 atmospheric deposition as revealed by sampling individual rain events in the region of Geneva, Switzerland. *J Environ Radioactiv* 53:241–256
65. Nijampurkar VN, Rao DK (1993) Polar fallout of radionuclides ^{32}Si , ^7Be and ^{210}Pb and past accumulation rate of ice at Indian station, Dakshin Gangotri, East Antarctica. *J Environ Radioactiv* 21:107–117
66. Kaste JM, Norton SA, Hess CT (2002) Environmental chemistry of beryllium-7. *Rev Mineral Geochem* 50:271–289
67. Lal D, Malhotra PK, Peters B (1958) On the production of radioisotopes in the atmosphere by cosmic radiation and their application to meteorology. *J Atmos Terr Phys* 12:306–328
68. Wallbrink PJ, Murray AS (1994) Fallout of ^7Be over south eastern Australia. *J Environ Radioactiv* 25:213–228
69. Harvey MJ, Matthews KM (1989) ^7Be deposition in a high rainfall area of New Zealand. *J Atmos Chem* 8:299–306
70. Turekian KK, Benninger LK, Dion EP (1983) ^7Be and ^{210}Pb total deposition fluxes at New Haven, Connecticut, and at Bermuda. *J Geophys Res* 88:5411–5415
71. Todd JF, Wong GRF, Olsen CR, Larsen IL (1989) Atmospheric depositional characteristics of Beryllium-7 and Lead 210 along the southeastern Virginia Coast. *J Geophys Res* 94:11106–11116
72. Peirson DH (1963) Beryllium-7 in air and rain. *J Geophys Res* 68:3831–3832
73. Lal D, Nijampurkar VN, Rajagopalan G, Somayajulu BLK (1979) Annual fallout of ^{32}Si , ^{210}Pb , ^{22}Na , ^{35}S and ^7Be in rains in India. *Proc Indian Acad Sci* 88A:29–40
74. Schumann G, Stoeppeler M (1963) Beryllium 7 in the atmosphere. *J Geophys Res* 68:3827–3830
75. Lee SC, Saleh AI, Babavali AD, Jonoby L, Kuroda PK (1985) - Beryllium-7 deposition at Fayetteville, Arkansas and excess polonium-210 from the 1980 eruption of Mount St Helens. *Geochem J* 19:317–322
76. Walton A, Fried RE (1962) The deposition of beryllium-7 and phosphorus-32 in precipitation at north temperate latitudes. *J Geophys Res* 67:5335–5340

77. Othman I, Al-Masri MS, Hassan M (1998) Fallout of ^7Be in Damascus city. *J Radioanal Nucl Chem* 238:187–191
78. Duenas C, Fernandez MC, Carretero J, Liger E, Canete S (2002) Atmospheric deposition of ^7Be at a coastal Mediterranean station. *J Geophys Res* 106:34059–34065
79. Lomenick TF, Tamura T (1965) Naturally occurring fixation of cesium-137 on sediments of lacustrine origin. *Soil Sci Soc Am Proc* 29:383–386
80. Squire HM, Middleton LJ (1966) Behavior of ^{137}Cs in soils and pastures – a long term experiment. *Radiat Bot* 6:413–423
81. Frissel MJ, Pennders R (1983) Models for the accumulation and migration of ^{90}Sr , ^{137}Cs , $^{239+240}\text{Pu}$ and ^{241}Am in the upper layer of soils. In: Coughtrey PJ, Bell JNB, Roberts TM (eds) *Ecological aspects of radionuclide release*. Blackwell, Oxford, pp 63–72
82. Jamsangton J, Parkpian P, Delaune RD, Jugsujinda A (2004) Retention of $^{137}\text{cesium}$ in acid sulphate soils of South Central Thailand. *Chem Ecol* 20:241–256
83. Romney EM, Childress JD (1965) Effects of beryllium in plants and soil. *Soil Sci* 100:210–217
84. He Q, Walling DE (1996) Interpreting particle size effects in the absorption of ^{137}Cs and unsupported ^{210}Pb by mineral soils and sediments. *J Environ Radioactiv* 30:117–137
85. Kaste JM, Norton SA, Fernandez IJ, Hess CT (1999) Delivery of cosmogenic beryllium-7 to forested ecosystems in Maine. *Geol Soc Am Abstr* 31:A305
86. Bettoli MG, Cantelli L, Degetto S, Tositti L, Tubertini O, Valcher S (1995) Preliminary investigations on ^7Be as a tracer in the study of environmental processes. *J Radioanal Nucl Chem* 190:137–147
87. Walling DE, Schuller P, Zhang Y, Iroumé A (2009) Extending the timescale for using beryllium-7 measurements to document soil redistribution by erosion. *Water Resour Res* 45:W02418. doi:10.1029/2008WR007143
88. Tiessen KHD, Li S, Lobb DA, Mehuys GR, Rees HW, Chow TL (2009) Using repeated measurements of ^{137}Cs and modelling to identify spatial patterns of tillage and water erosion within potato production in Atlantic Canada. *Geoderma* 153:104–118
89. Walling DE, Quine TA (1991) Calibration of caesium-137 measurements to provide quantitative erosion rate data. *Land Degrad Rehabil* 2:161–175
90. Walling DE, He Q (1999) Improved models for estimating soil erosion rates from cesium-137 measurements. *J Environ Qual* 28:611–622
91. Walling DE, He Q, Appleby PC (2002) Conversion models for use in soil-erosion, soil-redistribution, and sedimentation investigations. In: Zapata F (ed) *Handbook for the assessment of soil erosion and sedimentation using environmental radionuclides*. Kluwer, Dordrecht, pp 111–164
92. Loughran RJ, Elliott GL (1996) Rates of soil erosion in Australia determined by the caesium-137 technique: a national reconnaissance survey. International Association of Hydrological Sciences Publication No. 236. IAHS, Wallingford, pp 275–282
93. Mitchell JK, Bubenzer GD, McHenry JR, Ritchie JC (1980) Soil loss estimation from fallout cesium-137 measurements. In: DeBoodt M, Gabriels D (eds) *Assessment of erosion*. Wiley, Chichester, pp 393–401
94. Van Oost K, Govers G, Van Muysen W (2003) A process-based conversion model for caesium-137 derived erosion rates on agricultural land: an integrated spatial approach. *Earth Surf Process Land* 28:187–207
95. Quine TA, Walling DE, Govers G (1996) Simulation of radiocaesium redistribution on cultivated hillslopes using a mass-balance model: an aid to process interpretation and erosion rate estimation. In: Anderson MG, Brookes S (eds) *Advances in hillslope processes*. Wiley, Chichester, pp 561–588
96. Porto P, Walling DE, Ferro V (2001) Validating the use of caesium-137 measurements to estimate soil erosion rates in a small drainage basin in Calabria, southern Italy. *J Hydrol* 248:93–108
97. Sutherland RA (1992) Caesium-137 estimates of erosion in agricultural areas. *Hydrol Process* 6:215–225
98. Yan P, Dong Z, Dong G, Zhang X, Zhang Y (2001) Preliminary results of using ^{137}Cs to study wind erosion in the Qinghai-Tibet Plateau. *J Arid Environ* 47:443–452
99. Chappell A, Warren M (2003) Spatial scales of Cs-137-derived soil flux by wind in a 25 km² arable area of eastern England. *Catena* 52:209–234
100. Chappell A, Warren A, Oliver M, Charlton M (1998) The utility of ^{137}Cs for measuring soil redistribution rates in southwest Niger. *Geoderma* 81:313–337
101. Basher LR, Webb TH (1997) Wind erosion rates on terraces in the Mackenzie Basin. *J R Soc NZ* 27:499–512
102. Chappell A (1999) The limitations of using Cs-137 for estimating soil redistribution in semi-arid environments. *Geomorphology* 29:135–152
103. Schuller P, Walling DE, Sepulveda A, Trumper RE, Rouanet JL, Pino I, Castillo A (2004) Use of Cs-137 measurements to estimate changes in soil erosion rates associated with changes in soil management practices on cultivated land. *Appl Radiat Isot* 60:759–766
104. Schuller P, Walling DE, Sepulveda A, Castillo A, Pino I (2007) Changes in soil erosion associated with the shift from conventional tillage to a no-tillage system, documented using Cs-137 measurements. *Soil Tillage Res* 94:183–192
105. Walling DE, Zhang Y, He Q Models for deriving estimates of erosion and deposition rates from fallout radionuclide (caesium-137, excess lead-210 and beryllium-7) measurements and the development of user-friendly software for model implementation. In: IAEA TECDOC D1.50.08. IAEA, Vienna, (in press)
106. Schuller P, Iroumé A, Walling DE, Mancilla HB, Castillo A, Trumper RE (2006) Use of beryllium-7 to document soil redistribution following forest harvesting operations. *J Environ Qual* 35:1756–1763
107. Sepúlveda A, Schuller P, Walling DE (2008) Use of ^7Be to document erosion associated with a short period of extreme rainfall. *J Environ Radioactiv* 99:35–49

108. Olsson IU (1986) Radiometric dating. In: Berglund BE (ed) *Handbook of holocene palaeoecology and palaeohydrology*. Wiley, Chichester, pp 273–312
109. Appleby PG (2002) Chronostratigraphic techniques in recent sediments. In: Last WM, Smol JP (eds) *Tracing environmental change using lake sediments*, vol 1. Kluwer, Dordrecht, pp 171–203
110. Erten HN (1997) Radiochronology of lake sediments. *Pure Appl Chem* 69:71–76
111. Walling DE, He Q (1992) Interpretation of cesium-137 profile in lacustrine and other sediments - The role of catchment derived inputs. *Hydrobiologia* 235(236):219–230
112. Walling DE, He Q (1993) Toward improved interpretation of caesium-137 profile in lake sediments. In: McManus J, Duck R (eds) *Geomorphology and sedimentology of lakes and reservoirs*. Wiley, Chichester, pp 31–53
113. Walling DE (2010) Using fallout radionuclides to investigate erosion and sediment delivery: some recent advances. In: *Sediment dynamics for a changing future*, Proceedings of the Warsaw symposium, IAHS Publication no. 337. IAHS, Wallingford, pp 3–16
114. Golosov VN, Belyaev VR, Markelov MV, Kislenko KS (2010) Overbank sedimentation rates on the flood plains of small rivers in Central European Russia. In: *Sediment dynamics for a changing future*, Proceedings of the Warsaw symposium, IAHS Publication no. 337. IAHS, Wallingford, pp 129–136
115. Allison MA, Kuehl SA, Martin TC, Hassan A (1998) Importance of flood-plain sedimentation for river sediment budgets and terrigenous input to the oceans: insights from the Brahmaputra–Jamuna river. *Geology* 26:175–178
116. Walling DE, Owens PN, Leeks GJL (1998) The role of channel and floodplain storage in the suspended sediment budget of the River Ouse, Yorkshire, UK. *Geomorphology* 22:225–242
117. Walling DE, Owens PN, Leeks GJL (1999) Rates of contemporary overbank sedimentation and sediment storage on the floodplains of the main channel systems of the Yorkshire Ouse and the River Tweed, UK. *Hydrol Process* 13:993–1009
118. Walling DE, He Q (1997) Use of fallout ^{137}Cs in investigations of overbank sediment deposition on river floodplains. *Catena* 29:263–282
119. Hughes AO, Olley J, Croke JC, Webster IT (2009) Determining floodplain sedimentation rates using ^{137}Cs in a low fallout environment dominated by channel- and cultivation-derived sediment inputs, central Queensland, Australia. *J Environ Radioactiv* 100:858–865
120. Appleby PG, Oldfield F (1983) The assessment of ^{210}Pb data from sites with varying sediment accumulation rates. *Hydrobiologia* 103:29–35
121. Robbins JA (1978) Geochemistry and geophysical application of radioactive lead. In: Nriagu JO (ed) *The Biochemistry of lead in the environment*. Elsevier, Amsterdam, pp 285–393
122. He Q, Walling DE (1996) Use of fallout Pb-210 measurements to investigate longer-term rates and patterns of overbank sediment deposition on the floodplains of lowland rivers. *Earth Surf Process Land* 21:141–154
123. Lambert CP, Walling DE (1987) Floodplain sedimentation: a preliminary investigation of contemporary deposition within the lower reaches of the River Culm, Devon, UK. *Geogr Ann* 69A:393–404
124. Liu J, Carroll J, Lerche I (1991) A technique for disentangling temporal source and sediment variations from radioisotope measurements with depth. *Nucl Geophys* 5:31–45
125. Aalto R, Maurice-Bourgoin L, Dunne T, Montgomery DR, Nitttrouer CA, Guyot J-L (2003) Episodic sediment accumulation on Amazonian flood plains influenced by El Niño/southern oscillation. *Nature* 425:493–497
126. Aalto R, Dietrich W (2005) Sediment accumulation determined with ^{210}Pb geochronology for Strickland River flood plains. In: *Sediment budgets I*, Proceedings of the Foz do Iguaçu Symposium, Brazil, April 2005. International Association of Hydrological Sciences Publication No. 291. IAHS, Wallingford, pp 303–309
127. Canuel EA, Martens CS, Benninger LK (1980) Seasonal variations in the ^7Be activity in the sediments of Cape Lookout Bight. *Geochim Cosmochim Acta* 54:237–245
128. Dobb JE, Rice DL (1989) Temporal and spatial distribution of beryllium-7 in the sediments of Chesapeake Bay. *Estuar Coast Shelf Sci* 28:395–406
129. Fitzgerald SA, Klump JV, Swarzenski PW, Mackenzie RA, Richards KD (2001) Beryllium-7 as a tracer of short-term sediment deposition and resuspension in the Fox River, Wisconsin. *Environ Sci Technol* 35:300–305
130. Fisher GB, Magilligan FJ, Kaste JM, Nislow KH (2010) Constraining the timescales of sediment sequestration with large woody debris using cosmogenic ^7Be . *J Geophys Res* 115:F01013. doi:10.1029/2009JF001352
131. Walling DE (2005) Tracing suspended sediment sources in catchments and river systems. *Sci Total Environ* 344: 159–184
132. Collins AL, Walling DE, Leeks GJL (1998) The use of composite fingerprints to determine the provenance of the contemporary suspended sediment load transported by rivers. *Earth Surf Process Land* 23:31–52
133. Walling DE, Woodward JC (1995) Tracing sources of suspended sediment in river basins. *Mar Freshwater Res* 46:327–336
134. Walling DE, Owens PN, Leeks GJL (1999) Fingerprinting suspended sediment sources in the catchment of the River Ouse, Yorkshire, UK. *Hydrol Process* 13:955–975
135. Walling DE, Woodward JC (1992) Use of radiometric fingerprints to derive information on suspended sediment sources. International Association of Hydrological Sciences Publication No. 210. IAHS, Wallingford, pp 153–164
136. Collins AL, Walling DE, Leeks GJL (1997) Source type ascription for fluvial suspended sediment based on a quantitative composite fingerprinting technique. *Catena* 29:1–27
137. Collins AL, Walling DE, Webb L, King P (2010) Apportioning catchment scale sediment sources using a modified composite fingerprinting technique incorporating property weightings and prior information. *Geoderma* 155:249–261

138. Palmer MJ, Douglas GB (2008) A Bayesian statistical model for end member analysis of sediment geochemistry, incorporating spatial dependences. *Appl Stat* 57:313–327
139. Wallbrink PJ, Murray AS (1993) The use of fallout radionuclide as indicators of erosion processes. *Hydrol Process* 7:297–304
140. Yang M-Y, Walling DE, Tian J-L, Liu P-L (2006) Partitioning the contributions of sheet and rill erosion using Beryllium-7 and Cesium-137. *Soil Sci Soc Am J* 70:1579–1590
141. Whiting PJ, Bonniwell EC, Matisoff G (2001) Depth and areal extent of sheet and rill erosion based on radionuclides in soils and suspended sediment. *Geology* 29:1131–1134
142. He Q, Owens P (1995) Determination of suspended sediment provenance using caesium-137, unsupported lead-210 and radium-226: a numerical mixing model approach. In: Foster IDL, Gurnell AM, Webb BW (eds) *Sediment and water quality in river catchments*. Wiley, Chichester, pp 207–227
143. Whiting PJ, Matisoff G, Fornes W, Soster F (2005) Suspended sediment sources and transport distances in the Yellowstone basin. *Geol Soc Am Bull* 117:515–529
144. Olsen CR, Thein M, Larsen IL, Mulholland PJ, Cutshall NH, Byrd JT, Windom HL (1989) Plutonium, lead-210, and carbon isotopes in the Savannah estuary; Riverborne versus marine sources. *Environ Sci Technol* 23:1475–1481
145. Salant NL, Renshaw CE, Magilligan FJ, Kaste JM, Nislow KH, Heimsath AM (2006) The use of short-lived radionuclides to quantify transitional bed material transport in a regulated river. *Earth Surf Process Land* 32:509–524
146. Phillips JM, Russell MA, Walling DE (2000) Time-integrated sampling of fluvial suspended sediment: a simple methodology for small catchments. *Hydrol Process* 14:2589–2602
147. Ongley ED, Thomas RL (1989) Dewatering suspended solids by continuous-flow centrifugation: practical considerations. *Hydrol Process* 3:255–260
148. Owens PN, Walling DE (1996) Spatial variability of caesium-137 inventories at reference sites: an example from two contrasting sites in England and Zimbabwe. *Appl Radiat Isot* 47:699–707
149. Sutherland RA (1994) Spatial variability of ^{137}Cs and the influence of sampling on estimates of sediment redistribution. *Catena* 21:57–71
150. Sutherland RA (1996) Caesium-137 soil sampling and inventory variability in reference samples; literature survey. *Hydrol Process* 10:34–54
151. Campbell BL, Loughran RJ, Elliott GL (1988) A method for determining sediment budgets using caesium-137. International Association of Hydrological Sciences Publication No. 174. IAHS, Wallingford, pp 171–179
152. Zaborska A, Carroll J, Papucci C, Pempkowiak J (2007) Intercomparison of alpha and gamma spectrometry techniques used in ^{210}Pb geochronology. *J Environ Radioactiv* 93:38–50
153. Wallbrink PJ, Walling DE, He Q (2002) Radionuclide measurement using HPGe gamma spectrometry. In: Zapata P (ed) *Handbook for the assessment of soil erosion and sedimentation using environmental radionuclides*. Kluwer, Dordrecht, pp 67–96
154. Canberra (2010) Catalogue. <http://www.Canberra.com>
155. He Q, Walling DE (2000) Calibration of a field-portable gamma detector to obtain in situ measurements of the ^{137}Cs inventories in cultivated and floodplain sediments. *Appl Radiat Isot* 52:865–872
156. Benke RR, Kearfott KJ (2001) An improved in situ method for determining depth distributions of gamma-ray emitting radionuclides. *Nucl Instrum Meth Phys Res A* 463:393–412
157. Golosov VN, Walling DE, Kvasnikova EV, Stukin ED, Nikolaev AN, Panin AV (2000) Application of a field-portable scintillation detector for studying the distribution of ^{137}Cs inventories in a small basin in central Russia. *J Environ Radioactiv* 48:79–94
158. Walling DE, Zhang Y (2010) A national assessment of soil erosion based on caesium-137 measurements. *Adv Geocol* 41:89–97
159. Walling DE, Owens PN (2002) The role of floodplain sedimentation in catchment sediment and contaminant budgets. In: *The structure and management implications of fluvial sedimentary systems, Proceedings of the international symposium held at Alice Springs, September 2002*, IAHS Publication 276. IAHS, Wallingford, pp 407–416
160. Walling DE (2000) Linking land use, erosion and sediment yields in river basins. *Hydrobiologia* 410:223–240
161. Walling DE, Collins AL, McMellin GK (2003) A reconnaissance survey of the source of interstitial fine sediment recovered from salmonid spawning gravels in England and Wales. *Hydrobiologia* 497:91–108
162. Zhang X, Walling DE, Quine TA, Wen A (1997) Use of reservoir deposits and Cs-137 measurements to investigate the erosional response of a small drainage basin in the rolling loess plateau region of China. *Land Degrad Dev* 8:1–16
163. Walling DE (2006) Tracing versus monitoring: new challenges and opportunities in erosion and sediment delivery research. In: Owens PN, Collins AJ (eds) *Soil erosion and sediment redistribution in river catchments*. CABI, Wallingford, pp 13–27
164. Ketterer ME, Watson BR, Matisoff G, Wilson CG (2002) Rapid dating of recent aquatic sediments using Pu activities and $^{240}\text{Pu}/^{239}\text{Pu}$ as determined by quadrupole inductively coupled plasma mass spectrometry. *Environ Sci Technol* 36:1307–1311
165. Everett SE, Tims SG, Hancock GJ, Bartley R, Fifield LK (2008) Comparison of Pu and ^{137}Cs as tracers of soil and sediment transport in a terrestrial environment. *J Environ Radioactiv* 99:383–393
166. Tims SG, Everett SE, Fifield LK, Hancock GJ, Bartley R (2010) Plutonium as a tracer of soil and sediment movement in the Herbert River, Australia. *Nucl Instrum Meth Phys Res B* 268:1150–1154
167. Wu FC, Zheng J, Liao H, Yamada M (2010) Vertical distributions of plutonium and ^{137}Cs in lacustrine sediments in north-western China: quantifying sediment accumulation rates and source identifications. *Environ Sci Technol* 44:2911–2917

168. Kim CS, Lee MH, Kim CK, Kim KH (2008) ^{90}Sr , ^{137}Cs , $^{239,240}\text{Pu}$ and ^{238}Pu concentrations in surface soils of Korea. *J Environ Radioactiv* 40:75–88
169. Van Pelt RS, Ketterer M, Zobeck TM, Ritchie JC (2009) Anthropogenic radioisotopes to estimate rates of soil redistribution by wind [abstract]. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America. 1–5 Nov 2009. Pittsburgh. Abstract No. 2009.55438
170. Schimmack W, Auerswald K, Bunzl K (2002) Estimation of soil erosion and deposition rates at an agricultural site in Bavaria, Germany, as derived from fallout radiocesium and plutonium as tracers. *Naturwissenschaften* 89:43–46
171. Hancock GJ, Leslie C, Everett SE, Tims SG, Brunskill GJ, Haese R Plutonium as a chronomarker in Australian and New Zealand sediments: a comparison with ^{137}Cs . *J Environmental Radioactivity* (in press)
172. Amos KJ, Croke JC, Timmers H, Owens PN, Thompson C (2009) The application of caesium-137 measurements to investigate floodplain deposition in a large semi-arid catchment in Queensland, Australia: a low-fallout environment. *Earth Surf Process Land* 34:515–529
173. Schimmack W, Auerswald K, Bunzl K (2001) Can $^{239+240}\text{Pu}$ replace ^{137}Cs as an erosion tracer in agricultural landscapes contaminated with Chernobyl fallout? *J Environ Radioactiv* 53:41–57
174. Ketterer ME, Hafer KM, Mietelski JW (2004) Resolving Chernobyl vs global fallout contributions in soils from Poland using Plutonium atom ratios measured by inductively coupled plasma mass spectrometry. *J Environ Radioactiv* 73:183–201
175. Rawlins BG, Scheib C, Beamish D, Webster R, Tyler AN, Young ME (2010) Landscape-scale controls on the spatial distribution of caesium-137: a case study across Northern Ireland. *Earth Surf Process Landforms*. doi:10.1002/esp. 2026
176. Scheib C, Beamish D (2010) High spatial resolution observations of ^{137}Cs in northern Britain and Ireland from airborne geophysical survey. *J Environ Radioactiv* 101:670–680
177. Blake WH, Wallbrink PJ, Wilkinson S, Humphreys GS, Doerr SH, Shakesby RA, Tomkins K (2009) Deriving hillslope sediment budgets in wildfire-affected forests using fallout radionuclide tracers. *Geomorphology* 104:105–116
178. Zhang X, Walling DE (2005) Characterizing land surface erosion from cesium-137 profiles in lake and reservoir sediments. *J Environ Qual* 34:514–523
179. Owens PN, Walling DE, He Q, Shanahan J, Foster IDL (1997) The use of caesium-137 measurements to establish a sediment budget for the Start catchment, Devon, UK. *Hydrol Sci J* 42:405–423
180. Walling DE, Collins AL, Sickingabula HM, Leeks GJL (2001) Integrated assessment of catchment sediment budgets. *Land Degrad Dev* 12:387–415
181. Walling DE, Collins AL, Jones PA, Leeks GJL, Old G (2006) Establishing fine-grained sediment budgets for the Pang and Lambourn LOCAR catchments. *J Hydrol* 330:126–141
182. Walling DE, Collins AL (2008) The catchment sediment budget as a management tool. *Environ Sci Policy* 11:136–143

Books and Reviews

- Carroll J, Lerche I (2003) *Sedimentation processes: quantification using radionuclides*. Elsevier, Amsterdam
- Froehlich K (ed) (2010) *Environmental radionuclides: tracers and timers of terrestrial processes*. Elsevier, Amsterdam
- Garcia MH (ed) (2008) *Sedimentation engineering: processes, measurements, modelling and practice*. American Society of Civil Engineers, Reston
- Montgomery DR (2007) *Dirt: the erosion of civilisations*. University of California Press, Berkeley/Los Angeles
- Morgan RPC (2005) *Soil erosion and conservation*. Wiley-Blackwell, Oxford
- Owens PN (ed) (2008) *Sustainable management of sediment resources: sediment management at the river basin scale, vol 4*. Elsevier, Amsterdam
- Owens PN, Collins AJ (eds) (2006) *Soil erosion and sediment redistribution in river catchments*. CABI, Wallingford
- Zapata F, Nguyen M-L (2009) Soil erosion and sedimentation studies using environmental radionuclides. In: Froehlich K (ed) *Environmental radionuclides: tracers and timers of terrestrial processes*. Elsevier, Amsterdam

Fertilizer Science and Technology

J. F. ANGUS

CSIRO Plant Industry, Canberra, ACT, Australia

Article Outline

Glossary

Definition of the Subject

Introduction

Mining and Manufacture of Fertilizers

Fertilizer Distribution, Marketing, and Price

Fertilizer Effects on Agricultural Productivity

Fertilizer Recommendations and Decision Support

Fertilizers and Precision Agriculture

Fertilizers in Organic Farming

Environmental Benefits of Fertilizers

Pollution of Surface Water and Groundwater

Greenhouse Gas and Ammonia Emissions

Soil Acidification

Does N Fertilizer Deplete Soil Organic Matter (SOM)?

Nutrient Transfer from Farmland to Natural Terrestrial

Environments

Human Health

Resource Availability
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Essential nutrients There are 16 mineral elements that are essential for plant growth. They are referred to here by their symbols: nitrogen, N; phosphorus, P; potassium, K; calcium, Ca; magnesium, Mg; sulfur, S; boron, B; iron, Fe; manganese, Mn; copper, Cu; zinc, Zn; molybdenum, Mo; sodium, Na; chlorine, Cl; silicon, Si; nickel, Ni.

Fertilizer efficiency This refers to the additional grain or other agricultural product produced per unit of additional nutrient applied in fertilizer. A related concept is Fertilizer Recovery, which is the additional mass of a nutrient in the aboveground parts of a crop expressed as a proportion of the additional nutrient applied in fertilizer.

Macronutrients N, P, and K.

Micronutrients Cl, Fe, Mn, B, Zn, Cu, Mo, and Ni.

Soil microbial processes These include mineralization, which is oxidation of organic matter and the release of mineral nutrients; immobilization, which is the reverse process of incorporating mineral nutrients into organic matter; and nitrification, which is conversion of ammonium to nitrate.

Definition of the Subject

Fertilizers are compounds or mixtures delivered as solids, liquids or gases, that supply essential nutrients to crops in soluble forms that are convenient and safe to handle. Fertilizers may be applied to the soil or directly to foliage. All nutrients except N are manufactured by concentrating and refining ores extracted from mines. N fertilizers are manufactured from ammonia, which is synthesized from N_2 and H_2 . Science contributes to fertilizer use with improved products and methods to increase fertilizer efficiency, profitability of nutrient used, and reducing adverse environmental effects. Technology contributes to fertilizer use by improving the efficiency of manufacture and the complex logistics of delivering hundreds of million tons of products to farms safely, economically, and on time. Fertilizers can

be both inorganic and organic, but this contribution refers mostly to inorganic or manufactured fertilizers, since these provide most of the nutrients now added to soil and crops.

Introduction

Fertilizers are indispensable because nutrient supplies from the soil are normally inadequate for the high-yielding crops needed to supply food and fiber for the growing human population. Fertilizers replenish nutrients removed by previous crops and supplement the supply of nutrients from the soil to a level that will produce the farmer's target yield. Fertilizer may also improve the quality of human food, stockfeed, and fiber. The nutrients in fertilizer are a nonrenewable resource and, if they escape from the field, can cause environmental damage. This article discusses the sustainability of fertilizer in relation to these issues.

Before the invention of fertilizers, the only ways that farmers could accumulate nutrients on a field was to fallow before growing a crop, or to transfer animal manure from grasslands and barns to cropped land, either directly or by grazing livestock on grassland by day and enclosing them by night in a field prior to cropping. An additional source of N has been, and still is, the residues of previous legume crops, pastures, and green manures. With these natural nutrient sources, land is temporarily taken out of crop production. Fertilizer allows land to be used for continuous production.

The consumption of the macronutrients in fertilizer, N, P and, K has grown rapidly since the middle of the twentieth century and has increased the food supply more rapidly than the growth of the human population (Table 1). The increase in fertilizer supply has been even more rapid than the food supply. It is difficult to overstate the importance of fertilizers in sustaining humanity: Stewart et al. [1] concluded from the results of experiments that production of 30–50% of the world's food is a result of commercial fertilizers and Smil [2] calculated from nutrient balances that that nourishment of 40% of the world's population depended on N fertilizers.

The use of the macronutrients is uneven around the world, as shown in Fig. 1. On all continents, fertilizer provides relatively more N than P or K. The usage of all

Fertilizer Science and Technology. Table 1 Growth in world population, cereal production, and consumption of nutrients in fertilizer during the second half of the twentieth century

	Annual growth (%)
<i>Human population</i> ^a	1.8
<i>Cereals</i> ^a	
Maize	2.7
Rice	2.4
Wheat	2.3
<i>Fertilizer</i> ^b	
N	7.1
P	4.6
K	3.7

^a1961–2001 [3]

^b1960–1988 [4]

three nutrients in Europe and North America most closely approaches crop requirement while usage in Asia and Africa is relatively high in N. The K applications are negligible in Oceania and Africa probably because soil levels of K tend to be relatively high in dry regions.

This article discusses the sustainability of fertilizers from the perspectives of resources and logistics involved in supplying fertilizer to the farm, the on-farm effects on productivity, and on-site and off-site environmental effects.

Mining and Manufacture of Fertilizers

The raw materials for fertilizer manufacture, apart from N, are mined from ore deposits. In the century before 1920, most fertilizer N was also extracted from mines and, before that, from accumulations of nitrate in soil where farm animals had been housed for long periods. Much of this nitrate was used to manufacture gunpowder and other explosives as well as for fertilizer. The increasing demand for explosives and N fertilizer in the nineteenth century led to brutal mining industries that extracted guano (vitrified fecal deposits from sea birds, mostly on tropical islands) and nitrates, a sorry history told by Bown [7].

In modern mining industries, the nutrient compounds in the ore are concentrated by chemical or physical processes into fertilizers that are stable and available for plant uptake. In some cases, two or more nutrient elements are combined, by chemical reaction or physical mixing, into compound fertilizers. Most fertilizers are applied as solids, and single-element and compound fertilizers are normally manufactured as relatively homogeneous granules, which flow freely, minimize water absorption, produce little dust, and which can be transferred efficiently between storages and applied to the soil using machinery or by hand. Most nutrients are available as several different chemical compounds. Examples of these compounds and the manufacturing processes are described by Tisdale et al. [8].

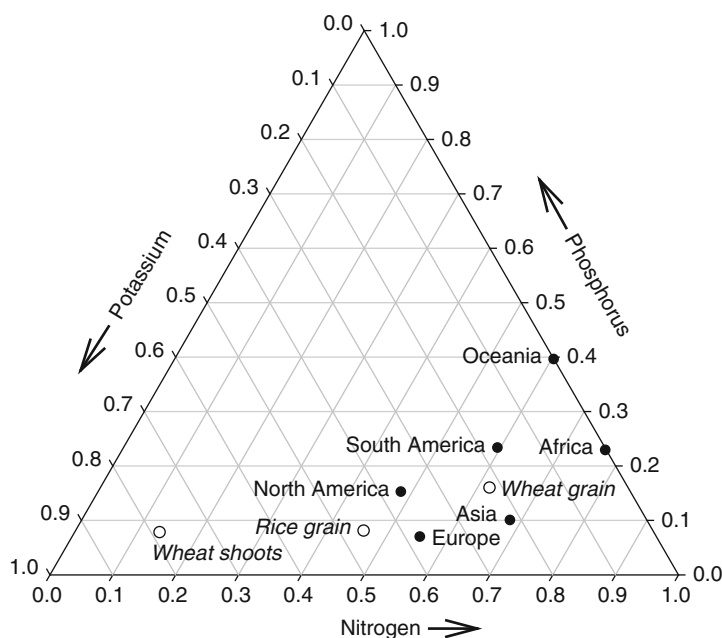
Nitrogen fertilizers are produced from ammonia, which is synthesized by the Haber-Bosch process under conditions of high temperature and pressure using N₂ and H₂ [2]. The usual source of H₂ is natural gas, and the price of N fertilizer is closely linked to the natural gas price (Fig. 2).

Many highly productive farming systems directly inject anhydrous ammonia into the soil. The advantage of ammonia is its high N concentration (82%) and hence low transport cost. Offsetting this is the need for expensive containment vessels, typically made from thick steel, and other costs of ensuring safety. The use of ammonia is confined to application at or before the time of sowing because of toxicity to growing plants. Applications of a large N amount at these times can cause yield reductions, as discussed below.

Ammonia is used to produce many forms of N fertilizer. The most common is urea, which has the advantages that it contains a relatively large percentage of N (46% N) and is safe to handle, unlike ammonium nitrate (34% N), which can be made to explode, and anhydrous ammonia, which is toxic to humans and animals at low concentrations in the atmosphere. Other common N-containing fertilizers are the solid ammonium phosphates (10–20% N) and the solution of urea-ammonium nitrate (28–32% N).

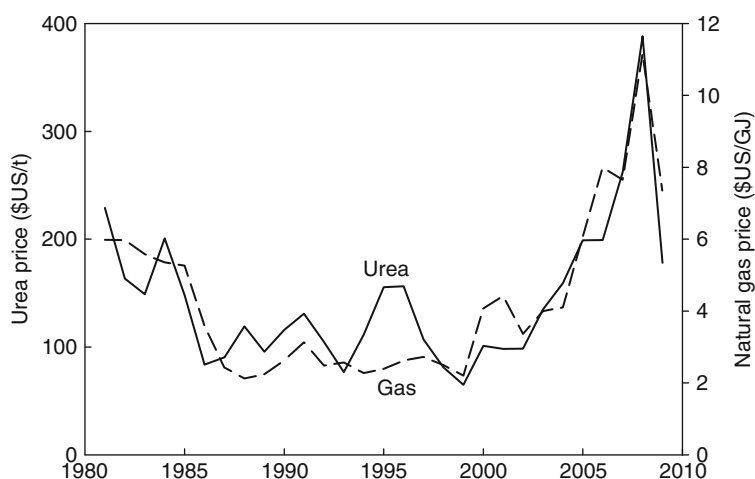
Fertilizer Distribution, Marketing, and Price

Most fertilizer in developing countries is granular and is sold in 40–50 kg bags and broadcast by hand, apart from



Fertilizer Science and Technology. Figure 1

Variation between continents in the proportions of N, P, and K used in fertilizer, and some examples of the proportions of these nutrients contained in whole crops (fertilizer data from FAOstats, rice data from Dobermann and Fairhurst [5], and wheat data from Bar-Tal et al. [6])



Fertilizer Science and Technology. Figure 2

The price of urea generally follows the price of natural gas, which represents the main production cost [9]

a small but increasing amount applied by machine at the time of sowing. In developed countries, most of the solid fertilizer is supplied in bulk and increasing amounts of liquid fertilizer are used because of the convenience of

pumping the product between vessels during transport and application. These advantages can compensate for the additional cost of transporting the water in which the nutrients are dissolved. Most of the advantages of

liquid are for on-farm handling, so an efficient system is for fertilizers to be delivered as solids from the manufacturer to the reseller who then dissolves the solids just before delivery to the farm. The advantages of liquid fertilizer (and gaseous anhydrous ammonia) are greatest for highly productive regions where farms are close to resellers and where it is possible to grow two or more crops each year and spread the cost of specialized equipment for mixing and applying fertilizer over a large crop area. Resellers extend the time of distribution by offering discounts for early use, but this can be at the expense of reduced fertilizer efficiency. One of the hazards of liquid fertilizer is soil compaction due to the mass of water and nutrients contained in equipment at the time of sowing or in spray carts for post-sowing applications.

Extensive dryland farming systems in developed countries use solid fertilizers where the transport costs are high. Solid fertilizer has an advantage for topdressing because granules can typically be mechanically thrown up to 10–15 m on each side of an implement, which is consistent with the 10–15-m width of seeders and 20–30-m width of boom sprays used for controlled traffic. The maximum spreading distance of granules may impose limitations as seeders and spray booms continue to get wider.

Some of the largest financial and environmental costs of fertilizer are related to transport. Products with low concentrations of nutrients are necessarily expensive to transport and their use is confined to specialized applications, for example calcium nitrate, which contains only 16% N, for some vegetables. During the second half of the twentieth century, there has been a general increase in nutrient concentrations leading to reduced cost of transport to the farm and handling on the farm. Nevertheless, the weight of fertilizer means that transport is expensive and low-cost transport methods such as ship, barge, and rail are preferable to road transport. Transport costs are minimized when fertilizer is back-loaded on vehicles carrying agricultural produce to cities. Transport makes up a large part of the on-farm cost of fertilizer in rural areas of sub-Saharan Africa and presents a major obstacle to increasing crop yields [10].

The transport cost is even greater for manures and compost than they are for inorganic fertilizer. Dry animal manures typically contain less than 2% N and 0.4% P, so

they cannot be economically carried far from the source. Instead, manures tend to be applied in a radius up to tens of kilometers. Where there is a large source of manure, such as a feedlot or large dairy, the available nutrients can overload the capacity for crop uptake, so that nutrients may accumulate in the soil within a few years. Where the manure contains potentially toxic material such as Cu contained in pig manure, nutrient loading of the soil needs to be closely monitored.

Fertilizer represents the main demand for N, P, and K, but for the micronutrients, usage for fertilizer manufacture is much less than for manufacturing industry. In the case of Zn, annual world production is about 20 million tons, but the amount of Zn contained in all cereal grains is less than 0.1 million tons per year.

The use of compound fertilizers has advantages in ease of handling and may be cheaper than single-element fertilizers. For example, the nutrients in ammonium phosphates are generally cheaper to the farmer than when they are purchased separately, for example as urea and triple superphosphate. In such situations, compound fertilizers can be profitable for the farmer, even if one of the nutrients is not particularly deficient. However with some NPK blends, there may be situations where crops do not respond to all nutrients contained in the fertilizer, and where the compound fertilizer is expensive, the whole product may be unprofitable to use. Such a situation applied to many farms in the Philippines where yield of rice did not respond to either P or K in a NPK compound fertilizer [11]. In this situation, application of the NPK was unprofitable on one third of farms and there was no simple way of determining which fields were deficient in which nutrient element. In such situations, the simplest solution is to make single-element fertilizers available so that strip trials conducted by farmers can show which nutrients give profitable responses [12].

The main fertilizers such as urea, MAP (monoammonium phosphate), and DAP (diammonium phosphate) can be regarded as generic and unspecialized commodities that are traded freely. They are sold by manufacturers to distributors on world markets that are generally open and transparent. An insight to the wholesale market is available online [13]. There is little opportunity for manufacturers or distributors to compete other than through price and service. Intense price competition

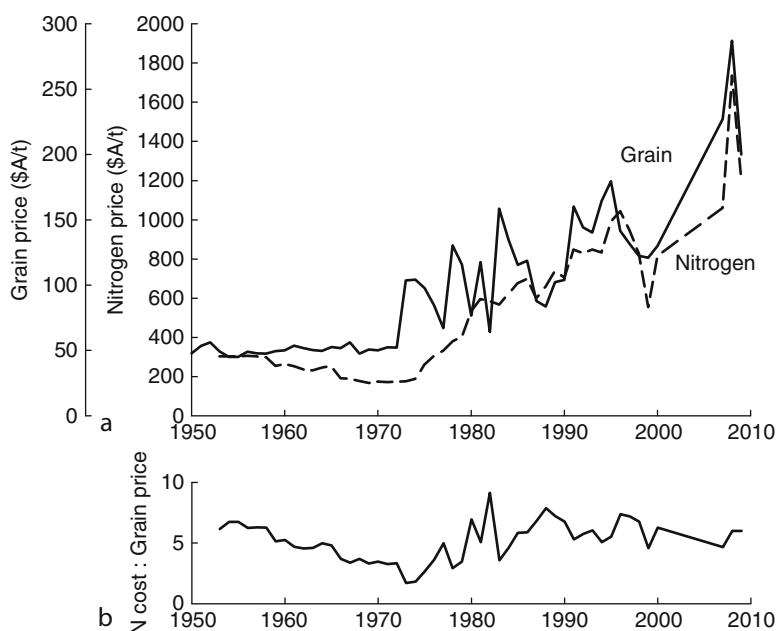
does not support the extension services offered by many fertilizer companies until the 1980s or 1990s.

Marketing is generally in the hands of agribusiness employees and media campaigns. Increasingly, advice about fertilizers is supplied to farmers at or near the point of sale by fee-for-service agronomists employed by fertilizer vendors. This represents a conflict of interest between the need for agribusiness to maximize sales and minimize carryover of fertilizer inventory, and the need of farmers to apply the fertilizer at an optimum amount and timing. Independent advice to farmers is important since fertilizer usually represents the single greatest cost of crop production.

Since generic fertilizers do not provide large profits, manufacturers or distributors aim to develop unique products that are easy to use or contain additives that increase fertilizer-use efficiency. Examples are additives to N fertilizers that inhibit urea hydrolysis and so limit the potential for ammonia volatilization, or nitrification inhibitors that restrict the production of nitrate and hence loss by leaching and denitrification [14]. The cost of these additives may not necessarily be justified by

additional profit, and they need to be assessed on a case-by-case basis [15]. Some urease and nitrification inhibitors are powerful biocides and have effects other than on N relations, for example, the nitrification inhibitor nitrapyrin has fungicidal activity, so any yield benefits may not be due only to effects on N supply to crops [16]. Slowing the release of nutrients from fertilizer by coating granules with a polymer is another possible means of maintaining adequate nutrient levels near the roots [17]. The cost-effectiveness of these products also needs to be assessed on a case-by-case basis. The most promising markets for fertilizer additives are where subsidies are offered to minimize N release into the environment.

World fertilizer prices fluctuate, depending on variations in supply and demand. The fluctuations for particular products can be massive, for example, the urea price spike in the mid-1990s when Chinese fertilizer companies suddenly increased urea imports (Fig. 2). The cost of fertilizers generally can also fluctuate strongly, as in 2008 when fertilizer price increased in tandem with a rise in world grain prices (Fig. 3).



Fertilizer Science and Technology. Figure 3

(a) Farm-gate cost of fertilizer N and price of wheat, both expressed in \$A and (b) the cost to price ratio for N and wheat grain (updated from Angus [18]). Costs and prices are expressed in Australian dollars (\$A) since these represent world prices undistorted by subsidies

It is clear from Fig. 3 that the price of grain and cost of fertilizer tend to readjust to maintain a reasonably constant cost:price ratio of about 6, even when grain price or fertilizer cost are disrupted by an external shock.

The price of fertilizer in many markets has been subsidized by governments, generally leading to fertilizer overuse, inefficient methods of application and, in some cases, excess agricultural production that is dumped on world markets. In some countries, the subsidy is through supply of natural gas at less than world price for ammonia production. Fertilizer subsidies have similar effects to subsidies on the price of agricultural production, in that they lead to overproduction because the optimum application rate of fertilizer is affected by the ratio of product price to fertilizer cost. Subsidies on agricultural products and fertilizers have long been criticized because they depress the price of agricultural products in world markets and so reduce the incomes of unsubsidized producers. Criticism is also justified because the overfertilization promoted by subsidies is a major cause of environmental damage [19, 20].

Fertilizer Effects on Agricultural Productivity

In their textbook on plant nutrition, Mengel and Kirkby [21] identify 16 mineral elements as essential for plants. Three of these, Na, Cl, and Si, are normally present in plant tissue at concentrations greater than needed for optimum growth and Ni is rarely or never included in fertilizers. Table 2 lists the remaining 12. A further two, cobalt and selenium, are not essential for plants but are essential animal (and human) nutrients and can be supplied in fertilizer or administered directly to farm animals. The other essential animal and human nutrient, iodine, is not normally supplied through fertilizer.

The concentration of the nutrients in plant tissue varies enormously; for example, N is 100,000 times more concentrated than Mo in cereal grain. The potential yield response to a nutrient, when it is deficient, can be estimated from the inverse of the nutrient concentration. So, for example, if the N concentration in grain is 0.02, as shown in Table 2, a first approximation of potential grain response is 50 kg grain per kg of additional N, assuming that the N concentration of the

Fertilizer Science and Technology. Table 2 Typical concentrations of nutrients in cereal grain, C, and the proportion of the aboveground nutrient in the grain [21], called the harvest index for the nutrient (HI). The maximum grain response to the nutrient is given by HI/C. The maximum net financial return is given by $[G(HI/C) - F]/F$ where F is the cost of nutrient in fertilizer and G is grain price, assumed here to be \$US 0.2/kg

Nutrient	Concentration in grain (C)	Proportion in grain (HI)	Maximum fertilizer efficiency (HI/C) (kg grain/kg fertilizer)	Nutrient cost (\$US/kg)	Maximum net returns (\$/\$)
N	0.02	0.7	35	1.0	6
P	0.003	0.7	233	5.0	9
K	0.004	0.2	50	0.8	12
Ca	0.005	0.7	140	0.2	140
Mg	0.003	0.5	167	1.1	30
S	0.003	0.4	133	0.2	130
B	0.0001	0.1	1,000	12	15
Fe	0.0002	0.3	1,500	3	100
Mn	0.00002	0.5	25,000	5	1,000
Cu	0.00002	0.5	25,000	5	1,000
Zn	0.00002	0.5	25,000	5	1,000
Mo	0.0000002	0.5	25,000,000	35	150,000

grain is unchanged by the fertilizer. In reality, not all the fertilizer N is present in the grain. In the case of wheat, 0.7 is a reasonable estimate of the proportion of above-ground N contained in mature grain. A better approximation of the maximum grain response to applied N is then $50 \times 0.7 = 35$ kg grain per kg N. Table 2 shows the equivalent calculation for all nutrients. This approach applies only to the first small application of a nutrient and does not account for the diminishing returns of yield to additional applications.

Table 2 also presents estimates of the maximum net financial return calculated from the maximum grain response, the cost of nutrients in fertilizer, and the price of grain. Note again that these are the maximum responses and returns, and real returns are normally much less because not all the fertilizer is recovered by crops. The costs and prices are necessarily assumptions and will vary greatly between markets and between years. However, the differences between nutrients are so great that even with approximations they serve to show some important points.

The maximum responses of additional grain per unit of nutrient applied in fertilizer in Table 2 can serve as benchmarks. When a crop is growing on soil that is believed to be deficient in a nutrient, it should give a yield response comparable with these benchmarks. If it does not it is likely that the nutrient is not seriously deficient or there is a limitation other than the deficiency that is limiting efficient utilization of the nutrient.

The maximum financial crop responses to the micronutrients such as Zn, Cu, and Mo are so large that they are normally applied if there is even a suspicion of deficiency. There are many farming systems where micronutrients still give major grain responses, for example in central Turkey, where widespread Zn deficiency was identified in the 1990s [22]. Continued vigilance is needed in all farming systems in case micronutrients become deficient.

Even for N and P fertilizer, the maximum financial responses are large enough to provide a profit even when the efficiency of recovery is low. In these examples crop recovery of only one sixth of the applied N or one ninth of the applied P gives a break-even return on cost of the fertilizer.

The grain-yield responses to fertilizers are usually much less than these maxima and a great deal of

research goes into increasing fertilizer efficiency. Where deficiencies are severe and there is only one nutrient lacking, visual symptoms can be compared with standard illustrations [23]. For less acute deficiencies, a traditional method is to calibrate yield against a soil test and identify the critical nutrient concentration that gives near-maximum yield [8]. The process of soil testing consists of sampling the soil in a field with many cores, combining the samples, and analyzing for nutrient concentration.

Soil tests are not the only methods for determining nutrient status and Table 3 shows examples of methods to estimate N status, including tests related to the nutrient status of vegetative plants. Tests of plant tissue are generally more reliable than soil tests because the volume of soil sampled by plant roots is normally more than the soil volume explored by coring [24, 25].

There are limitations to basing optimum fertilizer application only on tests of soil or plant nutrient status. The approach is suitable for a cropping system in which prices and yields do not vary greatly from year to year, but is less useful where the variation is large. The optimum fertilizer application varies with the ratio of grain price to fertilizer cost, and the optimum fertilizer rate is high when the ratio of grain price is high relative to fertilizer cost.

The optimum fertilizer rate also depends on yield potential, and higher rates are justified by increasing yield potential. Soil and plant tests emphasize the

Fertilizer Science and Technology. Table 3 Tests of N status

Pre-sowing tests	Tests during crop growth
Field history	Tissue N concentration
Total N in the topsoil	Tissue nitrate concentration
Mineral N in the topsoil	Sap nitrate concentration
N released during soil incubation	Shoot density
Previous grain protein	Leaf chlorophyll concentration
	Leaf color compared with standards
	Lower leaf senescence
	Strip trials

Fertilizer Science and Technology. Table 4 Example of a nutrient budget for N fertilizer applied to wheat

<i>Crop N demand</i>	
Target yield	4 t/ha
Target grain protein	12%
N in grain ^a	84 kg/ha
N in aboveground crop ^b	120 kg/ha
<i>Soil N supply</i>	
Soil mineral N at sowing ^c	50 kg/ha
N mineralization during crop growth	80 kg/ha
N available for crop uptake (50%) ^d	65 kg/ha
<i>Fertilizer N requirement</i>	
Additional N required in the crop ^e	55 kg/ha
Fertilizer N required ^d	110 kg/ha

^aGrain N is the product of yield and grain N, assuming that 1 kg of N = 5.7 kg of grain protein

^bAssuming grain N is 70% of aboveground N

^cSoil mineral N sampled to a depth of at least 60 cm

^dAssuming 50% of soil and fertilizer mineral N is available for crop uptake

^eDifference between crop demand and soil supply

supply of nutrients but it is important to place equal emphasis on crop nutrient demand. Table 4 shows an example of a nutrient budget, based on a method developed by Myers [26], to illustrate the importance of N supply and demand.

When the quality of the grain is affected by fertilizer, it is necessary to consider the responses of both yield and quality and the ratios of fertilizer cost to the grain price and quality premiums. Angus [27] presents a method to estimate the economic optimum fertilizer for wheat in relation to responses by grain yield and protein.

A well-known but generally erroneous notion about nutrients is the Law of the Minimum, which claims that only one nutrient can limit yield of a particular crop. Where one nutrient is drastically deficient, the Law of the Minimum is a fairly good rule of thumb, for example when a micronutrient such as Zn gives a spectacular yield response. More generally, this Law is deceptive, and crops and cropping systems tend to regulate the quantity of nutrients in the soil by luxury extraction of those that are adequate or surplus, and poor extraction of those that are deficient. The different rates of extraction continue until

many nutrients become co-limiting and are required for the yield to reach the biological potential [28].

The weakness of the Law of the Minimum can be seen from the many experiments where fertilization with any one of several nutrients can increase yield (e.g., [11]). Fertilizer management therefore usually needs to consider more than one nutrient. Where several nutrients simultaneously limit yield, the combined effect may be additive, meaning that the yield response of the combined nutrients is the same as the sum of yield responses of the nutrients applied singly. In other situations there may be a positive (or negative) interaction, meaning that the effect of the nutrients when applied together is greater (or less) than the sum of the effects of the individual nutrients. As with many aspects of fertilizer management, there is no alternative to conducting experiments on farms to quantify crop responses to nutrients and their interactions. This rest of this section discusses the macronutrients, N, P, and K, and, as examples of minor and micronutrients, S and Zn.

Nitrogen

Nitrogen is used in greater quantity than other nutrients but the financial returns are often less (Table 1). Nitrogen fertilizer has a reputation for low efficiency and many studies show that a crop typically recovers about half the fertilizer N applied, but the proportion can vary from zero to almost complete recovery [24]. The reasons for incomplete recovery are numerous but often the soil can supply as much N as the crop can take up, and any additional N remains in the soil where it may be lost in several processes. Nitrogen losses include leached nitrate, gaseous ammonia volatilization from urea and ammonia-based fertilizer, and nitrous oxide (N₂O) and N₂ from denitrification. Urea is at particular risk of loss, mainly because the soil becomes alkaline around a dissolving granule, so that any ammonium formed is rapidly converted to ammonia, which volatilizes.

Another cause of low crop recovery is immobilization of fertilizer N by soil microbes. This process is not necessarily a loss, since the N can be present in the soil organic matter (SOM) for one or more seasons, during which it is normally mineralized and taken up by later crops.

Nitrogen is unusual among the nutrients in that excess application often causes a reduction in crop

yield. In wet and high-yielding environments the reasons for yield reductions are lodging, because the crop becomes tall and top-heavy, and because of disease in thick foliage. In dry conditions, excess N stimulates growth, which can lead to exhaustion of soil water before maturity. Another and often more important reason is that high N levels reduce the concentration of water-soluble carbohydrates in the stems, which is an important source of assimilates for grain growth [29]. The time of application of N fertilizer is important in methods to avoid losses. Responses to topdressing N on wheat during the stem elongation phase are comparable to the responses to the same amount of N applied at sowing in both highly productive environments in Europe [30] and in water-limited environments in Australia [31]. Fertilizer N applied at sowing promotes rapid seedling growth, which leads to more risk of lodging, foliar disease, and haying-off than the same amount of N applied at stem elongation or later. Even when growing rapidly, small plants cannot take up large amounts of N, so mineral N in the soil at the time of sowing is at risk of loss from leaching and denitrification.

Mid-season application of N is effective because it coincides with the period of maximum crop uptake and so closely matches supply with demand. Application of N during or after the stem elongation phase in cereals usually increases grain protein more than application at sowing.

The N losses from irrigated rice can be particularly high. De Datta [32] reported that the average recovery of N fertilizer in rice experiments was 35–40%. Extensive research showed that the main pathways for N loss from lowland rice are ammonia volatilization and denitrification when urea is broadcast onto wet soil or shallow water. This research did not show ways to minimize losses and Fujisaka [33] suggested that research should focus more on methods to increase N efficiency rather than quantify the loss pathways. Subsequent agronomic research showed that with suitable methods of application the recovery of N, supplied as urea, can be as high as 75% for tropical rice [34] and 90% for temperate rice at a commercial scale [24].

The most efficient ways to increase N use efficiency (NUE) are different in each region and with each cropping system. Practices that are often effective are injecting fertilizer into the soil rather than broadcasting

it on the surface and delaying application of some or all of the fertilizer so that the supply of N is well synchronized with the crop demand. Other ways to increase NUE are to ensure that other factors are not limiting. These limitations can include deficiency of other nutrients, weeds, herbicide damage, and root disease. An example of the importance of overcoming such limitations is apparent in a 40% increase in dryland wheat yield in Australia during the 1990s, which coincided with the adoption of canola break crops and higher applications of N fertilizer. Angus [18] suggested that break crops controlled widespread root disease and provided suitable conditions for a large yield response to N fertilizer.

There have also been reports of improvements of NUE in a single season by classical plant breeding [35] or modification of single genes [36]. Genetic improvements in N use efficiency are promising but it is not yet clear whether they are consistent across environments and seasons, and whether they are as cost effective as the established and reliable improvements in crop management.

There is a widespread and recurring problem of farmers applying more than the economically optimum rate of N fertilizer. The problem was apparent in Europe and the USA from at least the 1970s [20] and the same pattern of overuse emerged in many other countries, including China since the 1980s [37]. N fertilizer applied in excess of crop requirements is sometimes called “insurance nitrogen” [38], implying that farmers are prepared to pay for the excess provided that the yield is maximized. Another explanation lies in the concept of “farming styles” of van der Ploeg [39], which interprets farming practices in relation to farmers’ goals. Farmers who want to boast of the highest district yields apply heavy inputs with little regard for input costs and net profit. These competitive individuals are uninterested in the profit-maximizing paradigm of agricultural science. To reverse the pattern of fertilizer overuse, it is important to redirect their competitiveness to efficiency rather than yield.

The reason that overfertilization is more common with N than other nutrients is not clear, but may be related to the rapid and visible response by crops. Within 5–10 days of N application, crops usually become greener and taller, which gratifies farmers, at least when they first observe the response. To this

extent, N fertilizer provides its own advertising. The greener and taller fertilized crops may or may not yield more than those that receive no additional fertilizer. Reasons for a lack of yield response may be that the real yield limitation is not N but some other limitation such as water deficit, disease, deficiency of another nutrient, frost, weeds, or herbicide damage. Since the 1990s, the application rates of N fertilizer have been static or falling in Europe and North America but crop yields have continued to rise [40]. Apparently, farmers and advisers in these regions are learning the lesson that the previous rates were excessive and the time of application was too early.

Phosphorus

Soil contains P as both soluble and insoluble inorganic phosphates and in organic forms including phytates, which make up more than 50% of the soil P. Fertilizer P adds to the soluble pool that can be taken up by crops, but the longer it remains in the soil the more is precipitated as insoluble iron and aluminum compounds in acid soil, or as insoluble calcium compounds in alkaline soil. Soils vary in the speed with which they precipitate soluble P, so soil tests are needed to measure plant-available P and the optimum rate of P fertilizer [8]. Tests of soil P and other nutrients based on ion exchange resins often give closer relations with plant uptake than other extraction methods [41]. Plant P tests are not as useful as plant N tests because P is applied at or before sowing.

Fertilizer management is designed to maximize and prolong P availability to the crop. The most powerful method is to include all the P in a band near the seed so as to maximize plant access before the P becomes unavailable due to precipitation. As with N fertilization, it is important to supply only enough for the crop since some of the excess P is likely to be precipitated. This applies whether the fertilizer is in the solid or liquid form. Liquid forms of P fertilizer are relatively more effective than solid forms on alkaline soils because the precipitated P is still partly available [42].

Most P in the soil is attached to clay particles and is not readily leached from any but the sandiest soils. Because most soil P is close to the surface it is unavailable when the topsoil dries out, so fertilizer injection below the surface can improve the efficiency

of uptake in dry environments [43]. Because of the advantages of banding and deep placement, all the P fertilizer needs to be applied at the time of sowing.

Potassium

Crops require a large supply of K but remove little in grain. Most of the K taken up by plants remains in the straw and returns to the soil when the straw decomposes or is burnt, provided the ash that remains after burning is not blown away in the wind. Once in the soil, K is not readily leached because of its positive charge and tends to be used by crops with greater efficiency than N or P. An exception is on soils that contain a large proportion of 2:1 clay minerals, where K is strongly fixed and becomes unavailable to roots.

Other Nutrients

The other nine nutrients commonly applied as fertilizers are not all discussed individually and the reader is referred to nutrient aspects by Mengel and Kirkby [21] and fertilizer aspects by Tisdale et al. [8]. This section discusses S and Zn because of their increasing importance.

Sulfur is present in the soil as sulfate, which as an anion is subject to leaching. Until the later part of the twentieth century, it was inadvertently applied in many regions. One form was in superphosphate, which was used for its P content (9%) but supplied S (11%) as well. The “high analysis” fertilizers, which largely replaced superphosphate, contain relatively small amounts of S. Another source of S for plants was atmospheric deposition as oxides of S from smokestacks, particularly on smelters of metal sulfides. This source of S is also decreasing because smokestacks are being fitted with “scrubbers” to reduce air pollution. Fertilizer S will be increasingly needed to replace these sources.

Zinc is important for not only increasing crop growth, but also as an essential human nutrient that is becoming seriously deficient. Zinc deficiency is widespread in soils and plants, particularly in west and south Asia [44], and methods and products are available in some countries to correct Zn deficiency [23]. Some, but not enough, compound fertilizers in developed countries include Zn but in many developing countries, Zn is available only as zinc sulfate.

Sometimes this expensive compound is diluted by unscrupulous retailers with salts that look similar, such as ammonium sulfate, so the product is an ineffective source of Zn (R.J. Buresh, 2011, personal communication).

Much of the Zn content of plants is transferred from the soil to plant roots by arbuscular mycorrhizal fungi (AMF). The large surface area of AMF filaments increases the ability of plants to explore the soil for immobile nutrients such as Zn. AMF are highly effective in soils with low nutrient status, but the modern farming practices of high levels of P fertilizer and growing brassica break crops tends to reduce AMF activity and increase the need for Zn fertilizer. Breeding plants for increased Zn uptake from soil can be effective in the short term but without supply of Zn fertilizer will exhaust soil reserves and hasten the onset of severe deficiency [45].

Fertilizer Recommendations and Decision Support

Methods for prescribing the fertilizer amount evolved from blanket recommendation provided by advisers, meaning the same practice irrespective of fields or seasons. Soil tests and nutrient budgets, as described earlier, give more differentiated prognoses of response can be used to develop “rules of thumb,” which are usually accepted by busy farmers. Computer-based decision support systems can take account of many factors that are known to affect nutrient response but are not as widely adopted as simple systems and rules of thumb. McCown [46] concluded that adoption of decision support systems had not met the promise they offered.

Fertilizers and Precision Agriculture

Precision agriculture is a set of principles that has developed since the 1990s, aimed at managing crops at a scale of several meters rather than as a whole field. It relies on a global positioning system (GPS) that estimates location with precision ranging from 4 to 0.02 m, spatial data about crop production such as maps from a yield monitor, biomass images from a satellite or aircraft, or spatial data about soil conditions such as electrical conductivity [47]. Fertilizer can

be managed by precision agriculture by varying the application rate according to the demand by the crop and/or the supply from the soil in different zones in a field. The farmer or adviser first prepares a prescription specifying the fertilizer rate for each zone and programs this in a computer carried in a tractor. This computer continuously determines where the tractor is located relative to the zones in the field, based on signals from the GPS receiver, and controls an actuator on the fertilizer implement that delivers the appropriate rate of fertilizer in each zone.

Yield maps provide useful data for deciding on fertilizer strategies and tactics. If one part of a field is consistently low yielding and it has previously received the same amount of fertilizer as the rest of the field, then it is likely that soil nutrient levels have accumulated. This pattern is common in the case of P fertilizer, and provides evidence to reduce the application rate on low-yielding zones. Variable applications of N fertilizer are a common use of yield maps. An example is in response to natural variation in soil organic matter, with additional N supplied to low-fertility zones [48]. Another example is in response to “unnatural” soil variation due to land leveling, where “cuts” expose subsoil containing low nutrient levels and “fills” consist of additional amounts of topsoil that is relatively fertile.

In these examples, additional fertilizer is designed to compensate for low soil nutrient status. If yield variation is due to soil properties other than nutrient status, then it may be inappropriate to compensate for low yield with more fertilizer and the best strategy is the exact opposite, that is, to apply more fertilizer to zones that consistently give the highest yields. An example of this situation is when spatial yield variation is due to the differences in the available soil water supply. Angus et al. [49] investigated yield and crop response to N across fields that varied in subsoil constraints of salinity, sodicity, and high concentration of boron. They showed that wheat yield and the crop response to N fertilizer were greatest on parts of a field with the least subsoil constraints, and yield responses to applied N were least on zones with the most severe constraints. Even in the absence of subsoil constraints, the soil water-holding capacity can control yield of dryland crops, and the highest rates of N fertilizer should be applied to zones with large water-holding capacity [50]. Spatial soil variation is often best measured by

electromagnetic induction, which directly measures apparent electrical conductivity, which is a proxy for other properties [51].

Where the response of a crop to fertilizer depends on its N status rather than other soil conditions, an optical sensor can provide useful information about a static or on-the-go estimate of crop N status. Optical systems using infrared sensors, such as the Greenseeker [52] and N-Sensor [53], can estimate N status of crops at a scale of about 1 m² while mounted on a tractor or fertilizer distributor, and can regulate the fertilizer rate on-the-go. A simple system using a domestic digital camera can provide equivalent information for a single scene [54].

The previous examples applied when the spatial yield patterns are consistent between years. A more complex approach is needed when the fertilizer response is inconsistent between zones and between years. For example in a dry season, N fertilizer may increase yield in low-lying, wet zones of a field but reduce yield in hilly, dry zones. Conversely, in a wet season, N fertilizer may be most effective in the hilly zones. The best response to this situation is not clear. One possibility is to examine yield maps for several years and manage fertilizer for an average season. Another is to delay the decision to apply fertilizer until the seasonal pattern is apparent. It may also be possible to minimize yield variation due to water movement by land-forming and drainage.

Fertilizers in Organic Farming

Many consumers are prepared to pay premiums for organically grown produce and a small proportion of farmers want to supply food produced by one of the organic farming systems. For certification of organic produce, the several organic farming systems generally require nutrients to be applied as manure and unprocessed minerals but not as processed fertilizer. Application of sufficient manure can provide optimum crop nutrition and until recently it was assumed that there were no disadvantages to the nutrient relations of organic farming. Research compiled by Kirchmann and Bergström [55] shows that this assumption is invalid because there is consistently more N leaching under organic than conventional systems, apparently because the soil N supply is not well synchronized with crop

N demand. These results suggest that the nutrient relations of organic farms may be less sustainable than conventional farms. Many governments subsidize organic farming systems and thus support unsustainable nutrient management.

In another comparison of organic and conventional farming systems, Fagerberg et al. [56] showed that nutrient levels of the organic part of a split farm remained adequate for several years after conversion from conventional, but that later some nutrients become depleted and productivity fell. The likely reason was that these nutrients had accumulated during several decades of fertilizer application before half the farm was converted to organic farming.

Residual nutrients from previous fertilizer application are not the only source of N in organic farming. In parts of north America and western Europe, the deposition of mineral N from the atmosphere contributes a large part of crop requirement [20]. The sources are ammonia emitted from manure in intensive animal industries and nitrogen oxides emitted from vehicles. Both sources of N originate from industrial operations, so it is difficult to understand how any farming conducted in such regions can be designated as organic.

Some strands of organic farming propose application of plant extracts and microbial preparations to the soil to promote crop growth, and such products have spread in the decades since there was regulation of agricultural chemicals. Reports of the effectiveness of some of these products were reviewed by Edmeades [57], who concluded that they gave no significant yield benefit when used as recommended. A manufacturer of one such product objected to Dr. Edmeades' comments and took legal action against him in New Zealand courts. Dr. Edmeades won the case but effectively lost his scientific career and wasted a year enmeshed in the legal system [58]. It is not clear how reputable science should deal with claims that "alternative" products provide production and environmental benefits out of all proportion to their nutrient composition. Even if one such product is proved to be ineffective and removed from the market, it is likely that others will take its place. A solution may be to strengthen consumer protection laws so that manufacturers can be penalized for making unsubstantiated claims.

Environmental Benefits of Fertilizers

Fertilizers provide environment benefit by increasing soil organic matter in nutrient-deficient soils. For example, superphosphate provided P and S for vast areas of deficient soil in Australia, which provided that trigger for biological nitrogen fixation by pasture legumes [59]. Even more spectacular increases in soil organic matter came when micronutrient deficiencies were corrected with fertilizers in the Western Australian sandplain. These are the only positive environmental effects of fertilizer and other effects are neutral or negative.

Pollution of Surface Water and Groundwater

When there is nitrate in the soil and the water supply from precipitation and irrigation exceeds evaporation, the nitrate will be leached and eventually enter the groundwater. Levels of groundwater nitrate began to increase in the second half of the twentieth century, most notably in Europe and North America, and have stubbornly remained high. For example, Johansson and Gustafson [60] reported that after N-fertilizer application to an arable field in southern Sweden was discontinued, there was a delay of 20 years before the concentration of subsoil nitrate decreased.

The best-known environmental damage from nutrients is water pollution in lakes, rivers, and parts of the ocean. The pollution causes hypoxia, algal blooms, and death of pelagic fish in many water bodies, the largest of which are the Gulf of Mexico [61], the Baltic Sea [62], and the Yellow Sea [63]. The delay in affecting nutrient levels in such water bodies is probably longer than for groundwater, as shown by the constant levels of dissolved N in the Baltic Sea 10 years after N fertilizer usage almost stopped in the eastern Baltic countries after the breakup of the Soviet bloc [62].

Both N and P from fertilizer contribute to pollution of large water bodies but even if N from fertilizer were not present, blue-green algae would be able to obtain their N requirements from biological fixation. Other evidence for the importance of P in water pollution is that the size of the “dead zone” in the Gulf of Mexico is more closely related to the discharge of P than of N [64]. The source of nonpoint P pollution may not be confined to fertilizer, and streambanks disturbed by humans and grazing animals have the potential to

release soluble P into water bodies. A promising method to reduce phosphate release into streams is to exclude grazing animals from riparian strips.

Systems to limit nutrient pollution of water have included fertilizer taxes and quantitative limits of nutrient loading from both fertilizer and manure. N fertilizer was taxed in Sweden for many years but there was little change in the rate of N application to crops. In the Netherlands, a system of nutrient accounting of inputs and outputs limits the nutrient balance of farms to reduce nutrient movement into groundwater and streams [65]. A promising method is to define sensitive zones, based on infiltration rate or proximity to streams, where fertilizers are banned, limited, or taxed [66].

Intensive research continues on the nitrogen cascade, a phrase that recognizes a chain of intended and unintended consequences when reactive N moves on land and in the air and water. An example of coordinated research in western Europe is given by Sutton et al. [67].

Greenhouse Gas and Ammonia Emissions

Nitrous oxide (N_2O) and N_2 are released from soil into the atmosphere by the processes of nitrification [68] and denitrification [69]. N_2O absorbs long-wave radiation and remains in the atmosphere for a long time so its greenhouse effect is much greater than that of CO_2 . The International Panel on Climate Change estimates the emission of N_2O from agriculture as a percentage of the N fertilizer applied. This approach takes no account of nitrification and denitrification of non-fertilizer sources, mainly from mineralization of organic N. Since fertilizers provide about half of the N used for food production, the contribution of the organic N to N_2O emission is probably underestimated. The most promising strategy to reduce N_2O loss are to minimize waterlogging, since denitrification proceeds most rapidly in anaerobic conditions.

Gaseous loss of NH_3 is greatest when ammonium is on or near the surface of an alkaline soil, or in the floodwater of rice. Windy conditions at the soil surface also contribute to rapid loss. In the worst conditions, the loss of fertilizer N can be up to 10–15% per day [15], but the loss can be reduced or eliminated by injecting ammonia-based fertilizers below the soil surface, or, if topdressing is unavoidable, by applying

fertilizer before forecast rain, so that the granules are dissolved and transported below the soil surface.

Soil Acidification

Ammonia-based fertilizers, including urea, acidify the soil when protons are released during nitrification. Soil pH falls more rapidly in light-textured and poorly buffered soils than in heavy-textured and highly buffered soils, but the process continues irrespective of soil type. The consequence of acidification is reduced plant growth, not because of the protons, but because of increasing concentrations of aluminum and manganese in the soil solution that accompany reduced pH. The most common cure is to apply lime, which represents a deferred and indirect cost of N fertilizers and an environmental cost since the reaction of lime on acid soil releases CO₂. A less common response is to apply nonacidifying fertilizers such as those containing nitrate, but these are more expensive than ammonia-based products.

Does N Fertilizer Deplete Soil Organic Matter (SOM)?

Until recently, there was a general belief that N fertilizer had little or no effect on SOM but this was challenged from analysis of soil data from the Morrow Plots, a long-term experiment at the University of Illinois in Urbana. Khan et al. [70] analyzed soil C and Mulvaney et al. [71] analyzed soil N from these plots and came to the startling conclusion that adding N fertilizer decreased SOM, apparently because the N fertilizer stimulated soil respiration. If true, this conclusion would prove that the modern N fertilizer industry is unsustainable.

Reid [72] and Powlson et al. [73] criticized these interpretations of the Morrow data. Their criticism was that the experiment had been confounded because the plot that received a large N dose from 1965 to 2005 had previously (1904–1966) received large annual doses of manure that had built up SOM to very high levels. After the manuring finished on that plot in 1967, the SOM decreased simply because it was above the equilibrium level for the environment, and not because of N fertilizer. This is a strong argument because there is ample evidence that soil maintains a high level of SOM

only if there is a constant input of organic matter to offset soil respiration, and this material contains C, N, P, and S at ratios close to that of SOM. Kirkby et al. [74] surveyed data on SOM content and estimated that the average C:N:P:S ratio of humus was 1000:80:20:14. After manuring stops, the SOM decreases for many decades until it returns to the equilibrium for the environment [73]. The decrease in SOM in the Morrow plots can be explained more by the oxidation of manure than the effect of N fertilizer. The specific criticisms by Reid [72] and Powlson [73] about the confounded treatments have not been challenged after 3 years of debate on the subject [75] so it appears that the concern about N fertilizer depressing SOM was a false alarm.

Nutrient Transfer from Farmland to Natural Terrestrial Environments

It has long been known that nutrients move from farmland to natural environments on land, as well as to freshwater and oceans, as described above. Pathways for nutrient transfer are water, gas, dust, and wild animals [76]. Movement by water is likely to remain in the stream system so is relatively unimportant for transfer to terrestrial environments. The only nutrient moved as gas is ammonia, and the main agricultural source of which is housed animals rather than emissions from fertilizer, although the original source of the nutrients is mostly fertilizer.

Nutrient movement in dust has not been widely discussed in the agricultural literature but is of more interest for earth science. Nutrients move in dust when wind lifts clay particles, which normally contain a large proportion of organic matter [77]. When the wind abates, the particles may land on another agricultural field where the nutrient transfer is relatively unimportant, and in natural environment where the nutrients may cause environmental damage.

The other source of nutrient transfer is through animals. Where farmland is adjacent to natural vegetation, it is common to see wild animals grazing on crops and pastures during the morning and evening and retreating into the cover of bush during the day. Examples are deer in Europe and North America, kangaroos in Australia, and birds almost everywhere. The obvious attraction to animals of the margin between farm and

bushland is access highly nutritious feed on the farms and cover in the natural vegetation. Nutrient transfer occurs through urine and dung deposited in the bushland. A consequence of nutrient enrichment of natural vegetation from all sources is eutrophication of the landscape, analogous to eutrophication of water [78], with weedy annual plants competing with native perennials. Nutrient transfer over short distances could place limits on the development of mosaic farming systems where farming and natural vegetation are distributed through a landscape according to land capability.

Human Health

Heavy metals can enter the human food chain from fertilizer. Cadmium (Cd) is the most toxic and is a contaminant in many phosphatic fertilizers [79]. When Cd is applied to a field it accumulates first in crops, grazing animals, and then in the tissue of people who consume the products. The most promising way of reducing Cd levels in fertilizer is by manufacturing phosphatic fertilizer from sources of rock phosphate that contain low levels of Cd [80].

Another supposed health hazard from fertilizer was nitrate, but the concern turned out to be a false alarm. From the 1940s until the 1990s, nitrate was believed to cause blue-baby syndrome (methemoglobinemia) because some cases of this disease that were associated with groundwater containing high nitrate levels, and this was the main reason for an official limit of 50 mg/L of nitrate in drinking water. Evidence reviewed by Addiscott [20] showed that nitrate at this concentration was harmless and that the original basis for this limit was faulty because water containing high nitrate concentrations also contained bacterial contamination, which was the real culprit for methemoglobinemia.

There are however beneficial effects of fertilizers on human health through micronutrients that are essential for human health. The best example is Zn, which is widely deficient in the human diet. The World Health Organization concluded that human Zn deficiency was one of the most serious causes of poor health [81]. Low levels and low bioavailability of Zn in human food are the cause. Increased use of zinc fertilizer would be a boon to human health, particularly to people who rely on a vegetarian diet, since plant products contain

less Zn than meat. In view of the difficulty in correcting Zn deficiency in crops, supplementation of human food with Zn may be more effective [45].

Resource Availability

World food supply relies on increasing amounts of nonrenewable resources to provide fertilizers. In the case of N, the resource involved is energy, and about 1% of the world's supply of fossil fuels is consumed in its production. Production of all the other nutrients depends on mining as well as relatively small amounts of energy used in manufacture and transport.

The energy cost of producing N fertilizer has decreased as technology improved. For example, the amount of natural gas needed to synthesize ammonia in the most efficient plants halved in the last 40 years of the twentieth century, and by 2000 was within about 25% of the highest possible efficiency [2]. With diminishing scope for improved efficiency and exhaustion of natural gas, it is likely that the real cost of N fertilizer will rise.

The nutrient at greatest risk of exhaustion is P. Cordell et al. [82] estimated that "peak phosphorus" had been reached and that reserves would be exhausted by 2030. This conclusion is based on the extrapolation of sigmoid curves of production, similar to those used to predict the exhaustion of crude oil reserves. Alternative approaches reviewed by Cornish [83] comes up with estimates of reserves lasting at least until 2100, based on current rates of extraction and the amounts and quality of known reserves compiled by the US Geological Survey. Whether the reserves will last for 30 or 100 years, there is increasing concern that P will be the most limiting nutrient in the long term. The most promising ways of prolonging supplies is to reduce overuse of P fertilizer and find ways to minimize phosphate fixation in soils. If exhaustion of P is as imminent as the worst predictions, it will be necessary to find ways of recycling at the important points of loss, which are in food preparation and in human urine. Recycling options are discussed by the Global Phosphorus Research Initiative [84].

Reserves of high-quality K fertilizers are likely to outlast those of P and probably N [83]. One concern about K supplies is that they are concentrated in the hands of relatively few producers who could be in

a position to corner the market. Another concern is that the largest reserves of K fertilizer in Canada and Russia are remote from the regions of greatest K deficiency in Asia.

Future Directions

Existing science and technology shows that crops can respond reliably and profitably to fertilizers when used cautiously. This involves applying fertilizer so that the combined supply from fertilizer and the soil is sufficient for, but does not exceed, crop requirement. This principle has not penetrated far into farming systems, and fertilizer management varies from generally inadequate applications in sub-Saharan Africa to excess use of N in East Asia. Many farms in Europe and North America have passed the period of excess N use. However the efficiency of fertilizer use is still generally low and improvement will require incremental research on farms to refine methods to apply fertilizer efficiently and in ways that are compatible with farming systems. As farming systems change, as they inevitably will, fertilizer management will have to change also. The most promising methods are to arrange fertilizer rate, time of application, spacing and depth of placement to synchronize supply with crop demand, and to adjust these parameters in relation to zones in the field and region. The methods will have to be cost effective, so it will be important to continue to develop methods to manage low-cost fertilizers appropriately rather than resorting to high-cost additives, which may have to be supported by subsidies. Science should also critically evaluate and comment on “alternative” fertilizer products and systems that make outrageous claims and will lead to reduced food security if widely used.

More fundamental science should be directed to liberating the large amounts of unavailable soil nutrients, while recognizing that release of these bound nutrients will only delay the need for additional fertilizer. The largest environmental improvements related to nutrient overuse are likely to come from research to increase fertilizer efficiency on farms. The delay between conduct of farm-related research and its benefits on a regional scale can be many decades [85], so well-resourced field research needs to be maintained.

Acknowledgments

Tony Good provided insights into fertilizer technology and John Passioura and Mark Peoples made helpful suggestions about the article.

Bibliography

Primary Literature

1. Stewart WM, Dobb DW, Johnston AE, Smyth TJ (1995) The contribution of commercial fertilizer nutrients to food production. *Agron J* 97:1–6
2. Smil V (2001) Enriching the earth. Fritz Haber, Carl Bosch, and the transformation of world food production. MIT Press, Cambridge
3. www.fao.org/waicent/portal/statistics_en.asp
4. Bumb BL (1995) World nitrogen supply and demand: an overview. In: Bacon PE (ed) Nitrogen fertilization in the environment. Marcel Dekker, New York, pp 1–40
5. Dobermann A, Fairhurst T (2000) Rice: nutrient disorders and nutrient management. PPI and IRRI, Singapore/Los Baños
6. Bar-Tal A, Yermiyahu U, Beraud J, Keinan M, Rosenberg R, Zohar D, Rosen V, Fine P (2004) Nitrogen, phosphorus, and potassium uptake by wheat and their distribution in soil following successive annual compost applications. *J Environ Qual* 33:1855–1865
7. Bown SR (2007) A most damnable invention. Dynamite, nitrates and the making of the modern world. Penguin, London
8. Tisdale SL, Nelson WL, Beaton JD, Havlin JL (1993) Soil fertility and fertilizers. Macmillan, New York
9. World Bank (2008) Agriculture for development. World Bank, Washington, DC
10. Sanchez PA (2002) Soil fertility and hunger in Africa. *Science* 295:2019–2020
11. Angus JF, Fazekas de St. Groth C, Tasic RC (1990) Between-farm variability in yield response to inputs of fertilizers and herbicide applied to rainfed lowland rice. *Agric Ecosyst Environ* 39:219–234
12. Angus JF, Marquez DA, Tasic RC (2004) Diagnosing variable nutrient deficiencies in rainfed lowland rice using strip trials. In: Fourth international crop science congress, Brisbane. http://www.crops science.org.au/icsc2004/poster/4/1/2/678_angusjf.htm
13. www.icis.com
14. Freney JR (1997) Strategies to reduce gaseous emissions of nitrogen from irrigated agriculture. *Nutr Cycl Agroecosyst* 48:155–160
15. Peoples MB, Freney JR, Mosier AR (1995) Minimizing gaseous losses of nitrogen. In: Bacon PE (ed) Nitrogen fertilization in the environment. Marcel Dekker, New York, pp 565–602
16. Martens DA, Bremner JM (1997) Inhibitory effects of fungicides on hydrolysis of urea and nitrification of urea nitrogen in soil. *Pesticide Sci* 49:344–352
17. Chen D, Freney JR, Rochester I, Constable GA, Mosier AR, Chalk PM (2007) Evaluation of a polyolefin coated urea

- (Meister) as a fertilizer for irrigated cotton. *Nutr Cycl Agroecosyst* 81:245–254
18. Angus JF (2001) Nitrogen supply and demand in Australian agriculture. *Aust J Exp Agric* 41:277–288
 19. Cottrell R (1987) *The sacred cow: the folly of Europe's food mountains*. Grafton, London
 20. Addiscott TM (2005) *Nitrate, agriculture and the environment*. CABI, Wallingford
 21. Mengel K, Kirkby EA (2001) *Principles of plant nutrition*. Kluwer, Dordrecht
 22. Cakmak I, Yilmaz A, Kalayci M, Ekiz H, Torun B, Erenoglu B, Braun HJ (1996) Zinc deficiency as a critical problem in wheat production in central Anatolia. *Plant Soil* 180:165–172
 23. Alloway BJ (2004) Zinc in soils and plant nutrition. International Zinc Association, Brussels, http://www.zinc-crops.org/library_general_zni_publications.html
 24. Russell CA, Dunn BW, Batten GD, Williams RL, Angus JF (2006) Soil tests to predict the response of rice to fertiliser nitrogen. *Field Crops Res* 9:286–301
 25. Webster R, Oliver MA (2007) *Geostatistics for environmental scientists*. Wiley, Chichester
 26. Myers RJK (1984) A simple model for estimating the nitrogen fertilizer requirement of a cereal crop. *Fertilizer Res* 5:95–108
 27. Angus JF (1995) Modelling N fertilization requirements for crops and pastures. In: Bacon PE (ed) *Nitrogen fertilization and the environment*. Marcel Dekker, New York, pp 109–127
 28. Sinclair TR, Park WI (1993) Inadequacy of the Liebig limiting-factor paradigm for explaining varying crop yields. *Agron J* 85:742–746
 29. van Herwaarden AF, Farquhar GD, Angus JF, Richards RA, Howe GN (1998) 'Haying-off', the negative grain yield response of dryland wheat to nitrogen fertiliser. I. Biomass, grain yield and water use. *Aust J Agric Res* 49:1067–1081
 30. Ellen J, Spiertz JHJ (1980) Effects of rate and timing of nitrogen dressings on grain yield formation of winter wheat (*T. aestivum* L.). *Fertilizer Res* 1:177–190
 31. Angus JF, Fischer RA (1991) Grain and protein responses to nitrogen applied to wheat growing on a red earth. *Aust J Agric Res* 42:735–746
 32. De Datta SK (1987) Advances in soil fertility research and nitrogen management for lowland rice. In: *Efficiency of nitrogen fertilizers for rice*. International Rice Research Institute, Los Baños, pp 27–41
 33. Fujisaka S (1993) Were farmers wrong in rejecting a recommendation? The case of nitrogen at transplanting for irrigated rice. *Agric Syst* 43:271–286
 34. Peng S, Cassman KG (1998) Upper thresholds of nitrogen uptake rates and associated N fertilizer efficiencies in irrigated rice. *Agron J* 90:178–185
 35. Mi G, Tang L, Zhang F (2000) Is nitrogen uptake after anthesis in wheat regulated by sink size? *Field Crops Res* 68:183–190
 36. Good AG, Johnson SJ, De Pauw M, Carroll RT, Savidov N, Vidmar J, Lu Z, Taylor G, Stroehrer V (2007) Engineering nitrogen use efficiency with alanine aminotransferase. *Can J Bot* 85:252–262
 37. Ju X, Liu X, Zhang F, Roelcke M (2004) Nitrogen fertilization, soil nitrate accumulation, and policy recommendations in several agricultural regions of China. *Ambio* 33:300–305
 38. Mitsch WJ, Day JW Jr, Gilliam JW, Groffman PM, Hey DL, Wang N (2001) Reducing nitrogen loading to the Gulf of Mexico from the Mississippi River Basin: strategies to counter a persistent ecological problem. *Bioscience* 51:373–388
 39. van der Ploeg JD (1994) Styles of farming: an introductory note on concepts and methodology. In: van der Ploeg JD, Long A (eds) *Born from within: practices and perspectives of endogenous rural development*. van Gorcum, Assen, pp 7–30
 40. Fischer RA, Byerlee D, Edmeades GO (2009) Can technology deliver on the yield challenge to 2050? FAO, Rome, <ftp://ftp.fao.org/docrep/fao/012/ak977e/ak977e00.pdf>
 41. van Raij B, Quaggio JA, da Silva NM (1986) Extraction of phosphorus, potassium, calcium, and magnesium from soils by an ion-exchange resin procedure. *Commun Soil Sci Plant Anal* 17:547–566
 42. Holloway RE, Bertrand I, Frischke AJ, Brace DM, McLaughlin MJ, Shepperd W (2001) Improving fertilizer efficiency on calcareous and alkaline soils with fluid sources of P, N and Zn. *Plant Soil* 236:209–219
 43. Cornish PS (1987) Effects of direct drilling on the phosphorus uptake and fertilizer requirements of wheat. *Aust J Agric Res* 38:775–790
 44. Sillanpää M (1990) Micronutrient assessment at the country level: an international study. FAO, Rome
 45. Ryan MH, McNerney JK, Record IR, Angus JF (2008) Zinc bio-availability in wheat grain in relation to phosphorus fertilizer, crop sequence and mycorrhizal fungi. *J Sci Food Agric* 88:1208–1216
 46. McCown RL (2002) Changing systems for reporting farmers' decisions: problems, paradigms and prospects. *Agric Syst* 74:179–220
 47. McBratney A, Whelan B, Ancev T, Bouma J (2005) Future directions of precision agriculture. *Precision Agric* 6:7–23
 48. Blackmer AM, White SE (1997) Using precision farming technologies to improve management of soil and fertiliser nitrogen. *Aust J Agric Res* 49:555–564
 49. Angus JF, Norton RM, Pedler JF, Walker CN (2004) Cereal response to N fertiliser in relation to subsoil limitations. In: *Proceedings of the fourth international crop science congress, Brisbane*. http://www.cropscience.org.au/icsc2004/poster/2/5/1/1170_angusjf.htm
 50. Wong MTF, Asseng S (2006) Determining the causes of spatial and temporal variability of wheat yields at sub-field scale using a new method of upscaling a crop model. *Plant Soil* 283:203–215
 51. Sudduth KA, Kitchen NR, Bollero GA, Bullock DG, Wiebold WJ (2003) Comparison of electromagnetic induction and direct sensing of soil electrical conductivity. *Agron J* 95:472–482
 52. Raun WR, Solie JB, Taylor RK, Arnall DB, Mack CJ, Edmonds DE (2008) Ramp calibration strip technology for determining midseason nitrogen rates in corn and wheat. *Agron J* 100:1088–1093

53. Reusch S, Link A, Lammel J (2002) Tractor-mounted multispectral scanner for remote field investigation. In: Roberts PC (ed) Sixth international conference on precision agriculture. ASA/CSSA/SSSA, Madison, pp 1385–1393
54. Li Y, Chen D, Walker CN, Angus JF (2010) Estimating the nitrogen status of crops using a digital camera. *Field Crops Res* 118:221–227
55. Kirchmann H, Bergström L (eds) (2008) Organic crop production – ambitions and limitations. Springer, Heidelberg
56. Fagerberg B, Salomon E, Jonsson S (1996) Comparisons between conventional and ecological farming systems at Öreby. *Nutrient flows and balance. Swed J Agric Res* 26:169–180
57. Edmeades DC (2002) The effects of liquid fertilisers derived from natural products on crop, pasture, and animal production: a review. *Aust J Agric Res* 53:965–976
58. Edmeades DC (2002) Science friction. The Maxicrop case and the aftermath. Fertiliser Information Services, Hamilton
59. Williams CH, Donald CM (1957) Changes in organic matter and pH in a podzolic soil as influenced by subterranean clover and superphosphate. *Aust J Agric Res* 8:179–189
60. Johansson G, Gustafson A (2004) Discharge and nutrient losses for the agrohydrological year 2002/03 and a long-term review. Swedish University of Agricultural Sciences, Division of Water Quality Management, Uppsala
61. Goolsby DA, Battaglin WA, Lawrence GB, Artz RS, Aulenbach BT, Hooper RP, Keeney DR, Stensland GJ (1999) Flux and sources of nutrients in the Mississippi-Atchafalaya River Basin. NOAA Coastal Ocean Program, Silver Spring, Maryland
62. Wulff FV, Rahm LA, Larsson P (2001) A systems analysis of the Baltic Sea. Springer, Berlin
63. Duan S, Zhang S, Huang H (2000) Transport of dissolved inorganic nitrogen from the major rivers to estuaries in China. *Nutr Cycl Agroecosyst* 57:13–22
64. Snyder CS (2008) Nutrients and hypoxia in the Gulf of Mexico – an update on progress, 2008. *Better Crops* 92:16–22
65. Ondersteijn CJM, Beldman ACG, Daatselaar CHG, Giesen GWJ, Huijse RBM (2002) The Dutch mineral accounting system and the European nitrate directive: implications of N and P management and farm performance agriculture. *Ecosyst Environ* 92:283–296
66. Wright S, Mallia C (2003) The potential for eco-taxes as instruments of sustainability. *J Transdiscipl Environ Stud* 2:1–14, www.journal-tes.dk
67. Sutton MA, Howard CM, Erisman JW, Billen G, Bleeker A, Grennfelt P, van Grinsven H, Grizzetti B (2011) The European nitrogen assessment. Cambridge University Press
68. Norton JM (2008) Nitrification in agricultural soils. In: Scheppers JS, Raun WR (eds) Nitrogen in agricultural systems. ASA/CSSA/SSSA, Madison, pp 173–199
69. Coyne MS (2008) Biological denitrification. In: Scheppers JS, Raun WR (eds) Nitrogen in agricultural systems. ASA/CSSA/SSSA, Madison, pp 201–253
70. Khan SA, Mulvaney RL, Ellsworth TR, Boast CW (2007) The myth of nitrogen fertilization for soil carbon sequestration. *J Environ Qual* 36:1821–1832
71. Mulvaney RL, Khan SA, Ellsworth TR (2009) Synthetic nitrogen fertilizers deplete soil nitrogen: a global dilemma for sustainable cereal production. *J Environ Qual* 38:2295–2314
72. Reid DK (2008) Comment on “The myth of nitrogen fertilization for soil carbon sequestration”. *J Environ Qual* 37:739
73. Powlson DS, Jenkinson DS, Johnston AE, Poulton PR, Glendinning MJ, Goulding KWT (2010) Comments on “Synthetic nitrogen fertilizers deplete soil nitrogen: A global dilemma for sustainable cereal production”. *J Environ Qual* 39:749–752
74. Kirkby CA, Kirkegaard JA, Richardson AE, Wade LJ, Blanchard C, Batten G (2011) Stable soil organic matter: a comparison of C:N:P ratios in Australian and other world soils. *Geoderma* 163:197–208
75. Mulvaney RL, Khan SA, Ellsworth TR (2010) Reply to comments on “Synthetic nitrogen fertilizers deplete soil nitrogen: A global dilemma for sustainable cereal production”. *J Environ Qual* 39:753–756
76. Likens GE, Bormann FH (1974) Linkages between terrestrial and aquatic ecosystems. *Bioscience* 24:447–456
77. Boon KF, Kiefert L, McTainsh GH (1998) Organic matter content of rural dusts in Australia. *Atmos Environ* 32:2817–2823
78. Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S (2002) Agricultural sustainability and intensive production practices. *Nature* 418:671–677
79. McLaughlin MJ, Tiller KG, Naidu R, Stevens DP (1996) Review: the behaviour and environmental impact of contaminants in fertilizers. *Aust J Soil Res* 34:1–54
80. Lægread M, Bøckman OC, Kaarstad O (1999) Agriculture, fertilizers and the environment. CAB International, Cambridge
81. WHO (2002) World health report. World Health Organization, Geneva, www.who.int/whr/2002
82. Cordell D, Drangert J-O, White S (2009) The story of phosphorus: global food security and food for thought. *Glob Environ Change* 19:292–305
83. Cornish P (2010) A postscript to “peak P” – an agronomist’s response to diminishing P reserves. Proc. 15th Australian Agronomy Conference. www.regional.org.au/au/pdf/asa/2010/7452_cornish.pdf
84. <http://phosphorusfutures.net/>
85. Alston JM, Andersen MA, James JS, Pardey PG (2010) Persistence pays: U.S. agricultural productivity growth and the benefits from public R&D spending. Springer, New York

Books and Reviews

- Bacon PE (ed) (1995) Nitrogen fertilization in the environment. Marcel Dekker, New York
- Hatfield JR (ed) (2005) The farmer’s decision. Soil and Water Conservation Society, Ankeny
- Marschner H (1995) Mineral nutrition of higher plants. Academic, London
- Mosier AR, Syers JK, Freney JR (eds) (2004) Agriculture and the nitrogen cycle. Island Press, Washington

Fission Reactor Physics

MICHAEL NATELSON

Retired from the Bettis Atomic Power Laboratory,
West Mifflin, PA, USA

Article Outline

Glossary
Definition of Subject
Introduction
Mass–Energy Relationship
Heavy Elements
Fission and Its Products
Cross Sections
Neutron Distributions
Fission Reactor Performance
Future Directions
Bibliography

Glossary

Fissile Fissile isotopes are fissionable by the capture of neutrons of any energy, but are especially easily fissioned by the capture of slow neutrons, for example, U^{233} , U^{235} , Pu^{239} , and Pu^{241} .

Fertile Fertile isotopes may be transmuted into fissile isotopes by neutron capture. The naturally occurring fertile isotopes are Th^{232} and U^{238} .

Critical A critical fission reactor is in a steady state, with its neutron population sustained by a chain reaction.

Reactivity Reactivity is a dimensionless parameter, which characterizes how far from critical a fission reactor is. If zero, the reactor is critical; if positive, the reactor is supercritical and its neutron population is increasing; if negative, the reactor is subcritical.

Microscopic cross section A microscopic cross section is a parameter, with dimensions of area, that is a measure of the probability of a particular reaction resulting from an incident particle on a target nucleus. The *macroscopic cross section* for this “particular” reaction is the microscopic cross section times the number density of the target nucleus.

Definition of Subject

At the end of the nineteenth century and through the first half of the twentieth century, revolutionary discoveries were made in physics, and the laws of physics and our understanding of them were greatly expanded. In addition, tragic historical events led to an unprecedented concentration of intellectual talent and economic resources (the Manhattan Project) that allowed the new physics to be applied to the engineering of nuclear (fission) reactors. This entry will describe the advances in physics, which are key to fission reactor design, and how they enable this engineering practice.

Introduction

In 1900, Lord Kelvin (William Thomson) reportedly told the British Association for the Advancement of Science that “there is nothing new to be discovered in physics now. All that remains is more and more precise measurements.” Whether he actually said this or not, it is reasonable to believe that many scientists and engineers of his day would have concurred. Newton’s definitions and laws of mechanics and optics had long been successfully applied. Maxwell’s equations, Ohm’s law, etc. seemed to describe electricity and magnetism. Boltzmann and Gibbs had provided the foundations of statistical mechanics and thermodynamics. And chemists had been busy developing atomic theory, identifying 92 elements, the laws of chemical combination, the weights and sizes of atoms and molecules, and the periodic system.

With hindsight it is clear, however, that in 1900 there were many intriguing questions outstanding in the physical sciences, and there was an historically large cohort of scientists, being produced by the major universities of the day, ready to address them. The questions (and their resolutions) of prime importance to “fission reactor physics” are:

1. Does a theory of relativity apply to Maxwell’s equations, and is there a unique frame of reference (ether) for the propagation of light?
2. Why are the heaviest naturally occurring elements unstable, giving off various forms of “radiation” and transmuting to different elements?
3. What does the quantization of electromagnetic radiation (required to describe black body

radiation energy spectra and the photoelectric effect) mean to the laws of physics on the atomic scale?

The resolution of each of these questions will be discussed in this entry, as they are the starting points for the accumulation of knowledge needed to characterize the workings of fission reactors.

Clearly, Einstein's Theory of Relativity addressing question (1), and its identification of mass as a form of energy (1905) would, excuse the bad pun, energize the whole effort. Already in 1914, H. G. Wells in his novel "The World Set Free" envisioned industrial atomic energy and atomic bombs used in a catastrophic world war.

At the end of the nineteenth century, electrochemists looking for heavy elements (heavier than lead and bismuth) found that "radiation" was given off by the materials they were investigating. Becquerel (1896) observed γ rays (penetrating electromagnetic radiation similar to x-rays) from uranium salts. The Curies (1898) observed α and β rays from polonium and radium. Rutherford showed that the positively charged α s were doubly ionized helium atoms. The β s are negatively charged electrons, the same particles as the cathode rays that Thomson characterized and named (1897). These "radiations" proved to be key tools for determining the structure of atoms. The α particle was shown by Rutherford (1911) and his coworkers to scatter from gold foil in a manner inconsistent with the atomic model of the day, Thomson's *raisins* (electrons) *in the pudding* (positive charge medium) model. To explain the α scattering results, an atom's positive charge and its mass, minus that of its electrons, needed to be concentrated in a small nucleus (radius $\sim 10^{-12}$ cm), with its electrons distributed over a much larger volume (radius $\sim 10^{-8}$ cm), that of the whole atom. Niels Bohr, inspired by Rutherford's work, took to determining the *distribution* of atomic electrons. His success, building off Question (3) above, led to quantum mechanics. A complete model for the atom, however, still required an explanation for the mass of the nucleus. Again bombardment of various atoms (elements) with α particles led to the answer. Chadwick (1932) proved that the "rays" produced by α s striking beryllium nuclei were neutral particles with

mass slightly greater than the hydrogen nucleus, the proton. These neutral particles are the neutrons that had been hypothesized by Rutherford 12 years earlier. Heisenberg (1932) produced a detailed model of the atomic nucleus where the mass number A is the total number of elementary particles, protons plus neutrons, making up a nucleus, and the nuclear charge is Z , the number of protons. Thus, there can be various *isotopes* for a given element, more than one A for a given Z .

The discovery of the neutron marked the start of furious activity, culminating in the operation of the first fission reactor only 10 years later. Leo Szilard in 1933 recognized that a neutral neutron with modest kinetic energy could penetrate an atomic nucleus and cause a reaction releasing nuclear (mass) energy, and if, as part of the "reaction," additional neutrons were produced, a chain reaction could result. Szilard produced a patent for a reactor based on this idea and assigned it to the British Government in 1936 (before fission was discovered). In 1934, Fermi was using neutron bombardment (with neutrons of various energies) to produce nuclear transformations in many elements. Of special interest was the production of transuranic elements, Z greater than 92. Fermi won the 1938 Nobel Prize for this work. However, unknown at the time, he had also fissioned uranium. This was determined by electrochemical analysis of the products of neutron bombardment of uranium by Hahn and Strassmann. Subsequently, the process was identified as *fission* by Meitner and Frisch. Bohr recognized that the ease with which low energy neutrons could cause fission of uranium was due to the existence of the naturally occurring, but low atom percent (0.72%), isotope ${}_{92}\text{U}^{235}$ [1] (Various notations have been used to designate a particular isotope, for example, for uranium with mass number (A) 235; ${}_{92}\text{U}^{235}$, U^{235} , and ${}^{235}_{92}\text{U}$. The latter is in common use today. For ease of composition and for consistency with most of the references used in this entry the older standard, A as a right superscript, is used.). He and Wheeler, from their Theory of Fission [2], also recognized that the not yet produced isotope ${}_{94}\text{Pu}^{239}$, would also be readily fissioned by slow neutrons [3]. This was in early 1939. Bohr still did not think production of a fission bomb to be feasible.

Leo Szilard was, however, not deterred. He persuaded his friend Albert Einstein to write President

Roosevelt (8/2/1939), urging government support of fission research and the stock piling of uranium. This ultimately led to the Manhattan Project. In 1940, Seaborg and McMillan synthesized the readily fissionable isotope of plutonium, ${}_{94}\text{Pu}^{239}$, which is produced by neutron capture in the dominant uranium isotope ${}_{92}\text{U}^{238}$. Wheeler credited Louis Turner [3] with pointing out that kilogram quantities of ${}_{94}\text{Pu}^{239}$ could be produced in a large fission chain reaction reactor. Fermi and Szilard [4] designed and built the prototype for such a reactor, a “pile” of graphite blocks containing an array of natural uranium pellets. It was constructed in a squash court under a grand stand of the University of Chicago’s Stagg Field, and went *critical* (sustained a chain reaction) on December 2, 1942. The Manhattan Project built large reactors of this type for weapons material production, and also successfully pursued means of enriching uranium in ${}_{92}\text{U}^{235}$. Enriched uranium allows more compact, higher power density, reactor designs.

The Manhattan Project brought together extraordinary scientific and engineering talent, and immense resources to produce the weapons that ended the Second World War. It also provided the foundation for all fission reactor development that has followed. The subsequent advances in “physics,” which have contributed to this development, are principally:

1. The full understanding of the interaction of neutrons with nuclei: scattering (elastic and inelastic), and capture (simple absorption, transmutation, and fission), including measuring the parameters that characterize the probabilities of these “interactions”
2. The formulation of methods to solve the neutron transport (Boltzmann) equation, which governs the behavior of the dilute “gas” of neutrons in a fission reactor

This entry will discuss the topics, pre- and post-Manhattan Project, which encompass the physics of fission reactors.

Mass–Energy Relationship

In his initial paper [5] on the theory of relativity, Einstein confronted the problem of guaranteeing that

the laws of electromagnetism (Maxwell’s equations) apply in all inertial reference frames, just as the laws of mechanics do. In an inertial reference frame, an object, which is at rest, remains at rest and an object traveling with a particular velocity will maintain that velocity. Einstein asserted that there is no preferred reference frame (like stationary *ether* in space, as postulated years earlier), and that the speed of light c , in vacuum, 2.998×10^8 m/s, is the same in all inertial reference frames. From these assertions, Einstein derived transformations for various variables in the laws of physics from one inertial reference frame to another. This solved the “electromagnetism” problem and provided a firm grounding (theory) for phenomena observed when velocities approach the speed of light. For examples of the latter, see Kaplan, “Nuclear Physics” on the charge-to-mass ratio of the electron as a function velocity, and Mermin, “It’s About Time,” on the half-life of unstable particles as a function of their velocity. Our interest here is specifically on the relationship between mass and energy resulting from the special (not applying to gravity) theory of relativity. What is meant by the ubiquitous formula.

$$E = Mc^2? \quad (1)$$

For application to fission, an inelastic collision between two particles will be treated for relativistic conditions. The approach presented by Mermin in “It’s About Time” will be used.

In an elastic collision, total momentum, $\mathbf{P} = \mathbf{p}_1 + \mathbf{p}_2$, mass, $M = m_1 + m_2$, and kinetic energy, $K = k_1 + k_2$ are all conserved, where the mass, m , is an inherent property of a particle and is a measure of how it resists a change in its velocity. In an inelastic collision, only total momentum, $\mathbf{P}$ needs to be conserved. It needs to be conserved, however, in all inertial frames of reference. For relativistic conditions, one defines a particle’s momentum (a vector [in bold face]) as

$$\mathbf{p} = m\mathbf{u}/(1 - \mathbf{u}^2/c^2)^{1/2}, \quad (2)$$

where $\mathbf{u}$ is the particle velocity. As is required for consistency between relativistic and nonrelativistic laws of mechanics, Eq. 2 is effectively the nonrelativistic definition of momentum for the particle speed, $u \ll c$. Now to find $\mathbf{p}'$, the particle momentum, in

a frame moving with velocity $\mathbf{v}$ relative to the frame in which the particle has velocity $\mathbf{u}$, one applies the relativistic translation law for velocities:

$$\mathbf{u}' = (\mathbf{u} - \mathbf{v}) / (1 - \mathbf{u}\mathbf{v}/c^2). \quad (3)$$

Substituting for $\mathbf{u}'$ in the expression for $\mathbf{p}'$ (Eq. 2 with $\mathbf{p}$ and $\mathbf{u}$ primed), one obtains the relativistic translation law for momentum

$$\mathbf{p}' = (\mathbf{p} - p^0 \mathbf{v}) / (1 - \mathbf{v}^2/c^2)^{1/2}, \quad (4)$$

where

$$p^0 = m / (1 - \mathbf{u}^2/c^2)^{1/2}. \quad (5)$$

Now, if total momentum is to be conserved in our two-particle inelastic collision in both the primed and unprimed frames, then $P^0 = p_1^0 + p_2^0$ must also be conserved. Again, using the relativistic translation law for velocities (Eq. 3) and the definition p^0 (Eq. 5), we find that

$$p^{0'} = (p^0 - \mathbf{p}\mathbf{v}/c^2) / (1 - \mathbf{v}^2/c^2)^{1/2}. \quad (6)$$

And so for the *total* quantities we want to be conserved we have

$$\mathbf{P}' = (\mathbf{P} - P^0 \mathbf{v}) / (1 - \mathbf{v}^2/c^2)^{1/2} \quad \text{and} \quad (7)$$

$$P^{0'} = (P^0 - \mathbf{P}\mathbf{v}/c^2) / (1 - \mathbf{v}^2/c^2)^{1/2}. \quad (8)$$

Examining these expressions, it is clear that if $\mathbf{P}$ and P^0 are not changed after an inelastic (or elastic) collision, then neither is $\mathbf{P}'$ and $P^{0'}$.

In the limit of the speed u being much smaller than c , the difference between p^0 and m , ($p^0 - m$), approaches $mu^2/2c^2$. This result leads to a definition of relativistic kinetic energy, k , for a particle

$$k = p^0 c^2 - mc^2, \quad (9)$$

which has the required property of reducing to the nonrelativistic form, $mu^2/2$, in the limit of u much smaller than c .

Returning to our two-particle inelastic collision, as $\mathbf{P}$ is conserved so is $P^0 c^2$ and thus from Eq. 9

$$\Delta M c^2 = \Delta K, \quad (10)$$

where ΔM is the change in the masses of the inputs and outputs of the collision participants, and ΔK is the

change in the kinetic energies of these “inputs and outputs.” Thus, Eq. 10 provides insight into the meaning of “ $E = Mc^2$ ” for the fission process. For $n + {}_{92}\text{U}^{235} \mapsto$ fission products + 202.7 MeV (the ΔK of Eq. 10 in unit of millions of electron volts) the percent change in mass can be estimated by dividing 202.7 MeV by the energy equivalents of the inputs (i.e., 236 amu, where 1 amu = 931.141 MeV). The result is $\sim 0.1\%$, which may not appear to be large until one makes a comparison with a chemical reaction. For example, $\text{O}_2 + \text{C} \rightarrow \text{CO}_2 + 4.1 \text{ eV}$. A similar calculation indicates a $1 \times 10^{-8}\%$ conversion of mass to kinetic energy. Since one could not measure such a small change in total input and output masses in chemical reactants, it is not surprising that the full impact of “ $E = Mc^2$ ” had to await demonstration in a nuclear reaction like fission. However, as will be discussed in the next three sections, the large energy release in fission, while conforming to Eq. 10, is due to the strength of the forces that hold a nucleus together and the charge repulsion forces that will accelerate two smaller nuclei as they are formed in the fissioning of a larger “parent” nucleus.

Heavy Elements

The “heavy elements” of particular importance to fission reactors are the radioactive nuclei, which are characterized by systematic chains of decay. In nature, there are three chains (series). In a given series, each nucleus has a *mass number*, A , governed by a simple formula with the variable the integer n (see Table 1), and is identified with its longest lived isotope, that is, Thorium, Uranium, and Actinium (U^{235} had not been discovered when the $4n + 3$ series was identified). These *longest half-lives* are not surprisingly comparable to the age of the earth, 4.5×10^9 years. Half-life is one of three related parameters of radioactive decay processes, $T_{1/2}$, λ , and τ . The fundamental equation of radioactive decay is

$$-dN(t)/dt = \lambda N(t), \quad (11)$$

where λ is the *decay constant*, and $N(t)$ is the number of decaying nuclei at time t . The solution of Eq. 11 is

$$N(t) = N(0)e^{-\lambda t}. \quad (12)$$

The time when an original inventory of decaying nuclei, $N(0)$, is halved is

$$T_{1/2} = \ln 2/\lambda = 0.693/\lambda. \quad (13)$$

And as the decay process is statistical the *mean life-time*, τ , of a decaying nucleus is

$$\tau = (1/N(0)) \int_0^{\infty} N(0)\lambda te^{-\lambda t} dt = 1/\lambda, \quad (14)$$

the reciprocal of the decay constant.

With the search for transuranic elements through the bombardment of the heaviest natural elements, primarily with neutrons, a fourth decay series was identified, the Neptunium ($A = 4n + 1$) series whose radioactive members are not found in *nature* (see Table 1).

Of the “heavy elements,” the isotope U^{235} is key to fission reactor design. It is the only naturally occurring isotope which readily fissions when bombarded with neutrons of all energies. While its atomic percent abundance, 0.72%, is small, it is large enough to support chain reactions in reactors where neutrons born in fission are slowed down (moderated) by graphite (carbon) or by heavy water (deuterium oxide). When Uranium is enriched in U^{235} ($\sim 3\text{--}5\%$), it can fuel reactor designs where ordinary water moderates fission neutrons (today’s pressurized water and boiling water reactors). Having U^{235} available as a reactor fuel makes it possible to exploit the two abundant *fertile* “heavy elements,” U^{238} and Th^{232} . The term “fertile” refers to the fact that when these elements absorb a neutron they can be transmuted to *fissile* isotopes (Pu^{239} and U^{233} respectively), which like U^{235} readily fission when bombarded by neutrons of all energies. The transmutation processes are shown in Fig. 1. It is important to note that only one neutron capture is required in each of these transmutations. In a reactor design, neutron

economy is the key to maintain a chain reaction and, as will be discussed in the section on [Future Directions](#), expending one neutron with a reasonable probability of obtaining an additional fissile nucleus is a winner.

The heavy element radioactive decay series are also important to safety in fission reactor design. Each of the decay processes, α and β^- emissions and associated γ s, is favorable to energy release. So any heavy elements, particularly transuranics, in a reactor’s fuel system will contribute to the *decay heat* load that must be dissipated when a reactor shuts down. As will be discussed in the next section, the major short-term contributors to decay heat are *fission products*. A power reactor that shuts down following a sustained run at full rating will initially produce $\sim 7\%$ of that rating from decay heat, even if the chain reaction and nearly all fissioning has ceased.

For a full discussion of the radioactive decay series and the particulars of α , β^- , β^+ and γ emission, see Kaplan, and Krane, “Introductory Nuclear Physics.”

Fission and Its Products

As noted in the “[Introduction](#),” fission was discovered accidentally during the search for transuranic elements. This work by Fermi and others was part of an extensive effort to understand the atomic nucleus and to duplicate the great success of quantum mechanics and the Pauli exclusion principle in providing a fully predictive *Theory* of atomic electron structure. A comparable theory for the nucleus has not been developed, but several models (e.g., shell and liquid drop) provide insight into the trends and correlations found in the data provided by the extensive experimentation performed on the nuclei of the various elements and their isotopes.

Fission Reactor Physics. Table 1 Heavy element decay series

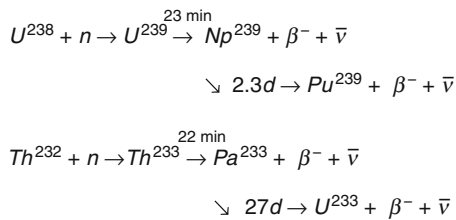
Series name	Type	Final stable nucleus	Longest lived nucleus	Longest half-life (years)
Thorium	$4n^a$	Pb^{208}	Th^{232}	1.41×10^{10}
Uranium	$4n + 2$	Pb^{206}	U^{238}	4.47×10^9
Actinium	$4n + 3$	Pb^{207}	U^{235}	7.04×10^8
Neptunium, not in <i>nature</i>	$4n + 1$	Bi^{209}	Np^{237}	2.14×10^6

^an is an integer

Measurements of atomic mass ($m(X^A)$), and the mass of the electron, proton, and neutron, yields the *binding energy*, B , of a nucleus, ${}_Z X^A$, the work (energy) required to disassemble a nucleus into its neutrons and protons:

$$B = \{Zm_p + Nm_n - (m(X^A) - Zm_e)\}c^2, \quad (15)$$

where Z is the atomic number (the number of protons) and $N = A - Z$ is the number on neutrons. (The binding

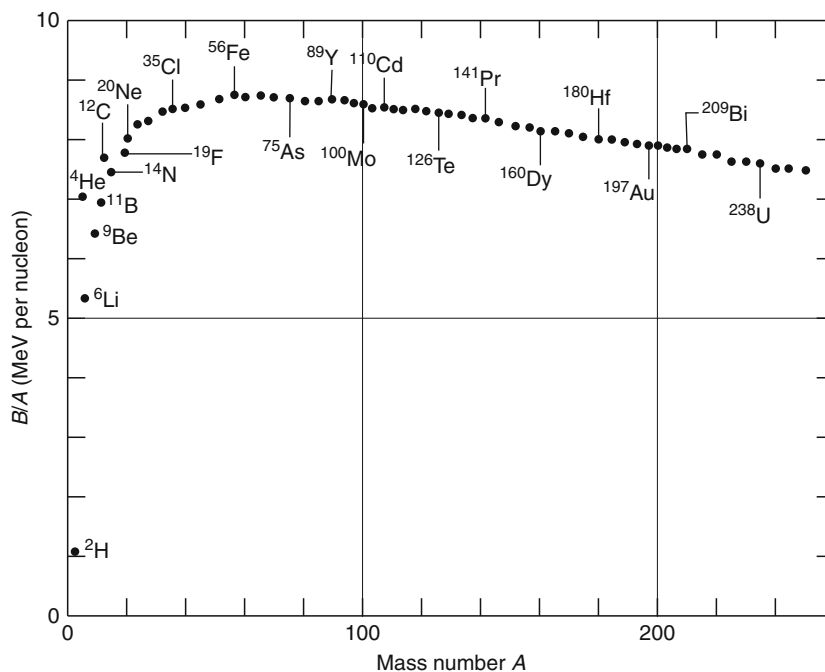


Fission Reactor Physics. Figure 1

Transmutation of fertile to fissile nuclei. ($\bar{\nu}$ is the antineutrino, the chargeless, $\sim$ zero mass particle that accompanies β^- emission)

energy of atomic electrons is ignored as negligible compared to the other factors in Eq. 15.)

Plotting the ratio of measured binding energies B to corresponding mass number A (Fig. 2) immediately makes evident the potential of energy release from fission of heavy element. Note the B/A versus A “curve” has a flat maximum in the middling A range $\sim 50 \rightarrow \sim 150$, and falls off (decreases) as A increases. Thus, there is a potential energy excess if a heavy element (isotope) can be disassembled and reassembled as two mid-range isotopes (preserving total A , Z , and N). (The behavior of the B/A curve for light elements shows the potential energy release from fusion.) Obviously fission (nor fusion) does not take occur “naturally” on earth today (There is convincing evidence that a naturally occurring chain reaction took place in a uranium deposit in Gabon about 2×10^9 years ago, when the abundance of U^{235} would have been $\sim 3\%$, high enough for a water-moderated “reactor” to operate. The higher earlier abundance is due to the shorter half-life of U^{235} (7.0×10^8 y) relative to that of U^{238} (4.4×10^9 y). See Krane for an excellent



Fission Reactor Physics. Figure 2

Binding energy per nucleon (Krane)

discussion of the Gabon reactor). The remainder of this section is devoted to particular requirements for fission to take place and to the discussion of the resulting fission products and their energies.

Insights provided by examining binding energies, and by additional experiments to determine nucleon–nucleon forces have led to the Shell and Liquid Drop models of the nucleus. Features of these models are incorporated in the *semiempirical mass formula* (Eq. 16). While a thorough discussion of the nuclear models is beyond the scope of this entry (see Kaplan or Krane), the mass formula provides key information on fission, energy release, and the relative likelihood for various nuclei.

In the semiempirical mass formula, the binding energy has five terms, which will be discussed below.

$$m({}_Z^AX^A) = Zm_p + Nm_n - [B_0 + B_1 + B_2 + B_3 + B_4]/c^2. \quad (16)$$

$B_0 = a_v A$ is the volume energy. Note in Fig. 2 that B/A saturates, thus B_0 has a linear dependence on A . The attractive nuclear forces between nucleons (n–n, n–p and p–p) are all equal and short range, smaller than the radius of the nucleus, $r = r_0 A^{1/3}$ where $r_0 \sim 1.2 \times 10^{-12}$ cm. If the range were larger, there would be attraction between each nucleon pair and B_0 would depend on $A(A-1)$.

$B_1 = -a_s A^{2/3}$, is the negative surface decrement. As the nucleon–nucleon forces are “short range,” neutrons and protons on the surface of a nucleus are less tightly bound.

$B_2 = -a_c Z(Z-1)/A^{1/3}$, is the coulomb repulsion decrement. While the nuclear forces are strong enough to overcome coulomb forces, the protons in the nucleus do repel and reduce binding energy. Assuming a uniform distribution of protons in a liquid drop model of a spherical nucleus, an electrostatics calculation yields the dependence of B_2 the number of proton pair, $Z(Z-1)$, and a measure of their spacing, $A^{1/3}$.

$B_3 = -a_a |N-Z||N-Z|/A$, is the neutron–proton population asymmetry decrement. As nuclei become heavier, more neutrons than protons are needed to overcome coulomb repulsion. However, as the shell model of the nucleus demonstrates when nucleons, neutrons and/or protons, are added to form heavier elements and their isotopes, they fill *shells* of

successively higher energy and are thus less tightly bound. This is analogous to the case of atomic electrons. Neutrons and protons have half-integral spin like the electrons, and therefore no two neutrons (or protons) can occupy the same state in a nucleus in conformance with the Pauli exclusion principle. So B_3 is negative and proportional to the neutron excess and the fraction of the nucleus the excess represents.

$B_4 = +\delta A^{-3/4}$ for even Z even N nuclei, $= 0$ for odd A nuclei, $= -\delta A^{-3/4}$ for odd Z odd N nuclei, is the pairing energy. As nucleons are added and fill *shells*, they are more tightly bound as spin up and spin down pairs. B_4 is important in determining the relative binding of isotopes of a given element and their propensity to fission.

A set of parameters for B which best fit the B/A curve (Fig. 2) is provided by Krane; $a_v = 15.5$ MeV, $a_s = 16.8$ MeV, $a_c = 0.72$ MeV, $a_a = 23$ MeV, and $\delta = 34$ MeV.

The potential for, and magnitude of, energy release from fission, whether as spontaneous decay or induced by particle or gamma ray capture, can be assessed with the semiempirical mass formula. As for an estimate of the magnitude of energy release, the B/A curve, as noted earlier, can be used directly. For example, the B/A for U^{238} is ~ 7.6 MeV. If it fissioned into two approximately equal mass nuclei ($A = 119$), their B/A would be ~ 8.5 MeV when in a ground state, and being more tightly bound than their parent (U^{238}) 214 MeV ($= 2 \times 119 \times 8.5 - 238 \times 7.6$) will be available through conservation of energy as kinetic energy of the daughter nuclei and of other fission products (neutrons, β s, γ s, and neutrinos). That this energy is *available* does not mean that there is a significant probability that fission occurs. In this example, which represents spontaneous fission of U^{238} , one finds in nature that this mode of U^{238} decay competes poorly with α decay (Spontaneous fission is a significant mode of decay for some transuranic isotopes found in depleted reactor fuel, particularly Pu^{240} and Pu^{241}). For fission fragments, daughter nuclei, to separate in spontaneous or induced fission, a potential barrier must be overcome. The height of the barrier relative to the ground state of a fission parent nucleus is called the *fission activation energy* (E_a). It can be estimated with the liquid drop model by calculating the change in the parent nucleus binding energy (B_1 and B_2) between

Fission Reactor Physics. Table 2 Heavy nuclei fission

Target nucleus	Compound nucleus	E_e , Excitation energy (MeV)	E_a , Activation energy (MeV)
U^{233}	$[U^{234}]$	6.6	4.6
U^{235}	$[U^{236}]$	6.4	5.3
Pu^{239}	$[Pu^{240}]$	6.4	4.0
U^{238}	$[U^{239}]$	4.9	5.5
Th^{232}	$[Th^{233}]$	5.1	6.5

the ground-state spherical configuration and a volume-conserving dumbbell configuration (ref. [2] and [6]). Table 2 contains values of E_a for the *compound* nuclei formed by neutron capture in the fissile and fertile isotopes of primary interest in reactor design. These are compared with the *excitation energy* (E_e) provided in forming the compound nucleus.

$$E_e = [(m({}_Z X^A) + m_n) - m({}_Z X^{A+1})]c^2. \quad (17)$$

Note that E_e does not include any kinetic energy contribution from the captured neutron. For the fertile target nuclei (U^{238} and Th^{232}), $E_e < E_a$ and neutron kinetic energy will be required to overcome or quantum mechanically penetrate (with high probability) the potential barrier to fission. For the fissile targets, $E_e > E_a$ and thus “slow” neutrons can initiate fission.

The high values of E_e for the fissile targets are due to the positive “pairing” contribution, B_4 , to the binding energy of the compound nucleus ground states. Note ${}_{92}U^{234}$, ${}_{92}U^{236}$, and ${}_{94}Pu^{240}$ are all even Z even N nuclei and the corresponding target nuclei are even Z odd N . So, the second term in Eq. 17 is decreased by $\delta(A + 1)^{-3/4}$, and B_4 is zero in the first term. Thus, an increase in E_e relative to the result if pairing is ignored is achieved. For fertile targets (even Z even N), roles are reversed. It is the first term in Eq. 17 that is decreased and B_4 is zero in the last term. Thus, E_e is lower than if pairing is ignored.

The semiempirical mass formula and the shell and liquid drop models are limited in predicting the fission process. This is best illustrated by the mass distribution of the major fission fragments (see Fig. 3). In the vast majority of cases, fission yields two unstable (having

excess neutrons) nuclei, but not of equal mass, as in the example above used to estimate the energy available from spontaneous fission of U^{238} . The two humped curves in Fig. 3 are not predicted by nuclear models. To quote Krane, “surprisingly, a convincing explanation for this mass distribution has not been found.”

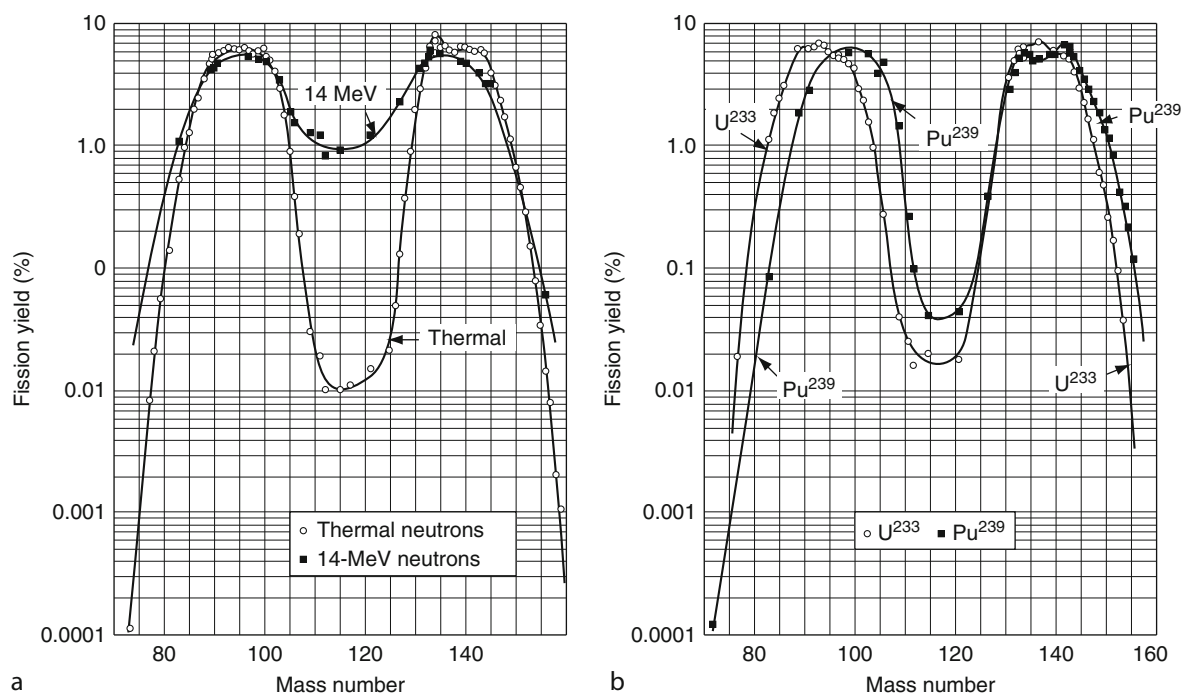
From the nuclear models, it is not surprising that free (prompt) neutrons are emitted in fission as the daughter nuclei are so rich in neutrons, but the prediction of their number (~ 2.5 on average) and energy spectrum (the mean ~ 2 MeV, see Fig. 4) are still an active area of study. The decay chains of the neutron-rich, excited daughter nuclei (fission fragments) are well predicted, including the release of (delayed) neutrons when in some cases neutron decay competes successfully with β -decay. The delayed neutrons are a small fraction of the total neutron emission (0.64% for thermal fission of U^{235}), but as will be discussed in section “Fission Reactor Performance”, they are important to reactor control.

Total energy release from the various neutron-induced fissions of interest in reactor design is remarkably consistent with the simple spontaneous U^{238} fission calculation made above. Of course, the constituents are different, as displayed in Table 3.

In a reactor design, the total energy values in Table 3 are not used. First, the contribution from neutrinos is subtracted, as their range before collision is well beyond reactor boundaries. Then, the energy release per fission from neutron captures which produce β s and γ s is added. The magnitude of this release is design-dependent as it is a function of the materials used, and the neutron capture rate in these materials. For plant energy balance studies, using 200 MeV/fission is satisfactory.

The problem of *decay heat* was noted in the previous section. From Table 3, it can be seen that fission product decay is the immediate concern when a chain reaction is terminated. Assume full power from U^{235} fissioning, when this ceases, delayed γ s and β s are still being released. Thus, $\sim 6.3\%$ ($\simeq 100 \times (6.26 + 6.43)/200$) of rated power, coming from fission product decay, must still be dissipated, along with energy from the decay of transuranic elements present in the reactor, for a total of $\sim 7\%$.

Over 800 fission fragment nuclei have been identified. Their decay must be tracked to account for *decay*



Fission Reactor Physics. Figure 3

Fission yields: (a) for U^{235} from fast and thermal neutrons, (b) for U^{233} and Pu^{239} from thermal neutrons [33]

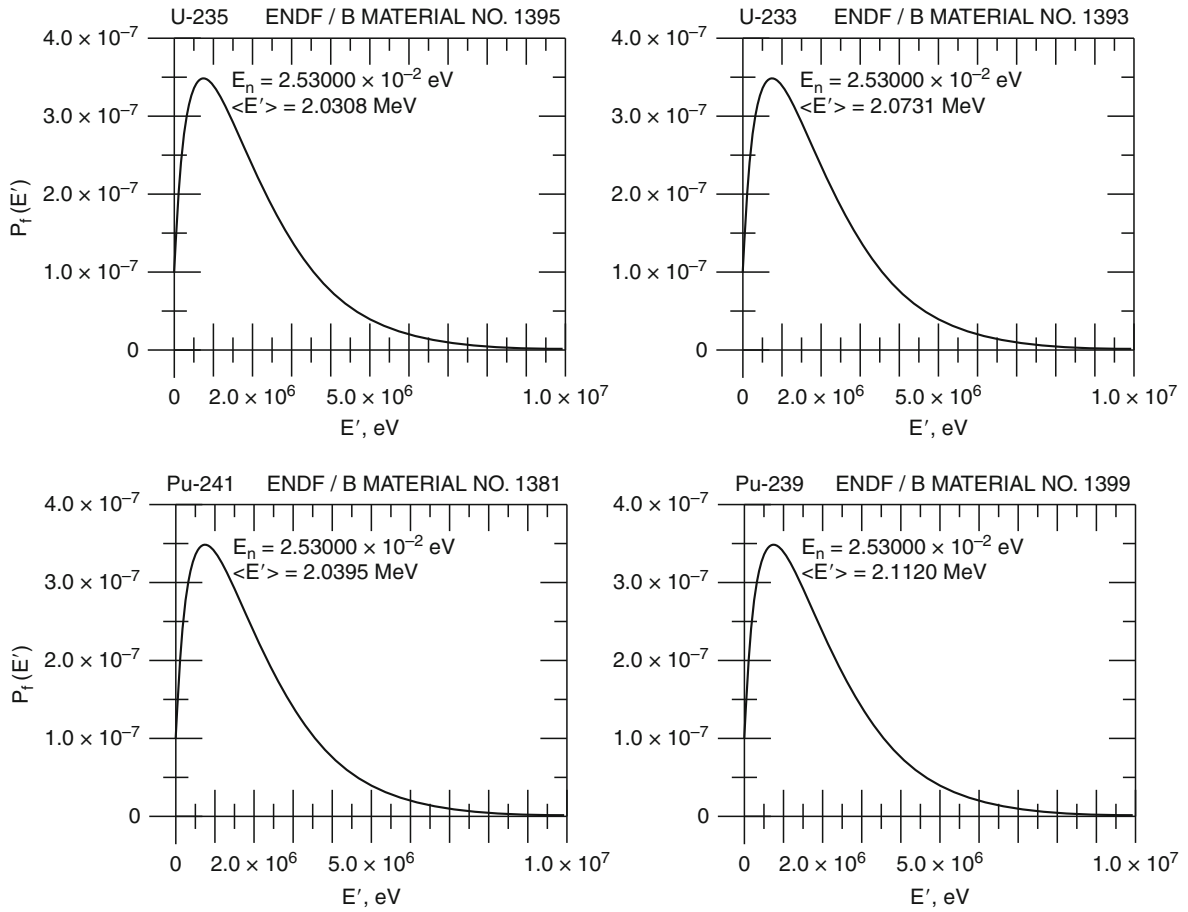
heat in reactor shut-down safety analysis and for the proper handling and storage of spent fuel (where both energy release and the nature of radiation fields must be known). One hundred and two of these nuclei are delayed neutron precursors. To simplify reactor transient (kinetics) calculations, the precursors are collected into six *effective* groups, where members of a given group have similar decay constants (see Table 4).

The energy spectra for a given delayed group do not vary significantly with fissioning isotope. The spectra are much softer (with lower mean energies, < 1 MeV) than for prompt neutrons [7]. This means that a delayed neutron in a thermal reactor is more *important* than a prompt neutron. It is more likely to reach the low energies (< 0.625 eV) where most fission occurs. Delayed neutrons can also result from other reactions, for example, photon capture (γ, n) (The expression (a,b) is shorthand for a nuclear reaction with an input particle “a” and output particle “b”, where the target and product nuclei are understood.) and neutron activation (n,p) followed by neutron decay of the product

nucleus. If important to a particular reactor design, these delayed neutrons can be included by modifying the effective delayed group structure.

The final aspect of the ~ 800 fission fragment nuclei that must be dealt with is their impact on neutron balance. Each of them has a probability of capturing neutrons and in some case of causing a transmutation into a nucleus with a particularly large propensity for capturing neutrons. The nuclei of greatest importance to neutron balance are listed in Table 5.

I^{135} is important as it is the direct precursor of Xe^{135} , an especially large absorber of thermal neutrons. The next two isotopes in the table are precursors to a decay chain with three large absorbers, Pm^{147} , Sm^{149} , and Sm^{151} . The final five, with their precursors in parentheses, are large absorbers, but not as sensitive to neutron energy spectrum and power level and history as the others. Clearly, data for the 800 fission fragments must be handled through large computer files [8]. For neutron balance, the fission fragment isotopes, which are not treated explicitly (Table 5), can be lumped into an effective fission product nucleus with



Fission Reactor Physics. Figure 4

Prompt neutron energy spectra where $P_f(E')$ is the probability per unit energy [32]

a yield per fission and *probability* for neutron capture. How one characterizes the *probability* of nuclear reactions is the subject of the next section.

Cross Sections

The nuclear reactions of importance to fission reactor design are by *definition* governed by the postulates of quantum mechanics (i.e., they are on the dimensional scale of the nucleus). And, thus the results of the various reactions are probabilistic in nature. The probability of a particular result is characterized by a parameter, the *microscopic* cross section, σ , with, not surprisingly, the dimensions of area, and which is quoted in units of *barns*. The barn, 10^{-24} cm^2 , is a reasonable measure as in some cases σ is nearly the

projected area of a target nucleus, $4\pi R^2$, and R is $\sim 10^{-12} \text{ cm}$. Thus, envisioning a target foil of area, A , and thickness, dx (where dx is small enough to have negligible shadowing of one nucleus by another in the target foil), the probability that an incident particle in traveling a short distance (i.e., dx) will undergo a specific reaction equals $\sigma \rho A dx / A$, where ρ is the density of target nuclei ($\#/\text{cm}^3$) in the foil. It follows that to find the reaction rate in the foil we need the number of impinging particles per second. Given the particles have a density N ($\#/\text{cm}^3$) and are monoenergetic and monodirectional (normal to the face of the foil) with speed v , the number impinging per second equals $N \cdot (vdt) \cdot A/dt$. So the total reaction rate in the foil is $(NvA)(\rho \sigma dx)$, and the rate per cm^3 is $vN\rho\sigma$. The parameters that make up this specific rate

Fission Reactor Physics. Table 3 Energy yield in MeV from fission. Thermal neutron fission of fissile isotopes; fertile isotope fission is from neutrons with energy spectrum of a light water-moderated reactor [7, 32]

Energy	Fissile or fertile isotope [†]					
	²³² Th	²³⁸ U	²³⁵ U	²³³ U	²³⁹ Pu	²⁴¹ Pu
Fission fragment kinetic energy	162.1 ± 1.5	170.0 ± 0.66	169.6 ± 0.7	168.7 ± 0.7	175.9 ± 0.1	175.5 ± 1.1
Prompt neutron kinetic energy	4.7 ± 0.12	5.51 ± 0.1	4.79 ± 0.07	4.9 ± 0.1	5.9 ± 0.1	5.99 ± 0.13
Delayed neutron kinetic energy	0.024 ± 0.004	0.021 ± 0.003	0.0071 ± 0.0007	0.0014 ± 0.0005	0.001 ± 0.0005	0.0074 ± 0.0015
Prompt γ rays	6.15 ± 1.75	6.28 ± 0.8	6.96 ± 0.7	7.59 ± 0.71	7.74 ± 0.45	7.86 ± 1.8
Delayed γ rays	8.01 ± 0.2	8.04 ± 0.08	6.26 ± 0.05	4.99 ± 0.04	5.16 ± 0.1	6.33 ± 0.07
β particles	8.28 ± 0.21	8.25 ± 0.08	6.43 ± 0.05	5.13 ± 0.04	5.3 ± 0.1	6.51 ± 0.04
Neutrino energy	11.1 ± 0.3	11.14 ± 0.11	8.68 ± 0.06	6.91 ± 0.05	7.15 ± 0.11	8.78 ± 0.09
Total energy yield	200.3 ± 0.5	209.3 ± 0.3	202.7 ± 0.1	198.2 ± 0.1	207.2 ± 0.3	211.0 ± 0.3

Fission Reactor Physics. Table 4 Prompt and delayed neutron data. ν_p is the prompt neutron yield versus initiating neutron energy (ENDF/B-VII.0). ν_d is the total delayed yield, constant for fission initiating neutron energies between 0 and 4 MeV, and above 7 MeV. Between 4 and 7 MeV, ν_d tracks linearly [32]

The relative abundance $\gamma_i = \nu_i/\nu_d$								
Neutron E		$^{233}\text{U } \nu_p$	$^{235}\text{U } \nu_p$		$^{239}\text{Pu } \nu_p$		$^{241}\text{Pu } \nu_p$	
0.0253 eV		2.4894	2.4208		2.8724		2.9251	
2 MeV		2.5758	2.6366		3.171		3.1727	
		$\bar{\nu}_d$						
E(MeV)		^{233}U	^{235}U		^{239}Pu		^{241}Pu	
0–4		0.0074	0.0167		0.00645		0.0162	
7		0.0047	0.0090		0.0043		0.0084	
^{233}U			^{235}U		^{239}Pu		^{241}Pu	
Delayed Group	Decay Constant $\lambda_{ir} \text{ s}^{-1}$	Relative Abundance γ_i	$\lambda_{ir} \text{ s}^{-1}$	γ_i	$\lambda_{ir} \text{ s}^{-1}$	γ_i	$\lambda_{ir} \text{ s}^{-1}$	γ_i
1	0.01258	0.086	0.01272	0.038	0.0129	0.038	0.0128	0.010
2	0.03342	0.274	0.03174	0.213	0.0311	0.280	0.0299	0.229
3	0.131	0.227	0.116	0.188	0.134	0.216	0.124	0.173
4	0.303	0.317	0.311	0.407	0.332	0.328	0.352	0.390
5	1.27	0.073	1.4	0.128	1.26	0.103	1.61	0.182
6	3.14	0.023	3.87	0.026	3.21	0.035	3.47	0.016

Fission Reactor Physics. Table 5 Direct yield fractions ($\times 100$) for isotopes in the most important fission product chains [32]

Fission Product	Fissile or fertile isotope ^a				
	²³² Th	²³⁸ U	²³⁵ U	²³³ U	²³⁹ U
¹³⁵ I	5.238	6.548	6.349	4.860	6.303
¹³⁵ Xe	0.0403	0.0150	0.255	1.337	1.152
¹⁴⁷ Nd	3.08	2.711	2.271	1.775	2.073
¹⁴⁹ Pm	0.825	1.765	1.089	0.769	1.261
⁹⁹ Mo (⁹⁹ Tc)	2.965	6.247	6.127	4.957	6.144
¹⁰³ Ru (¹⁰³ Rh)	0.164	6.336	3.137	1.707	6.991
¹³¹ I (¹³¹ Xe)	1.481	2.982	2.473	2.352	3.093
¹³³ I (¹³³ Cs)	3.858	6.356	6.787	5.974	6.923
¹⁴³ Ce (¹⁴³ Nd)	6.619	4.834	5.972	5.881	4.561

^aThe energy of the neutron initiating the fission is in the thermal range for ²³⁵U, ²³³U, and ²³⁹Pu. For ²³²Th and ²³⁸U, the yields are due to fissions initiated by neutrons with a spectrum of energies typical of light water-moderated nuclear reactors.

have been reordered to reflect conventional definitions in reactor physics (In the nuclear engineering discipline, *reactor physics* refers to the portion of the field addressed in this entry):

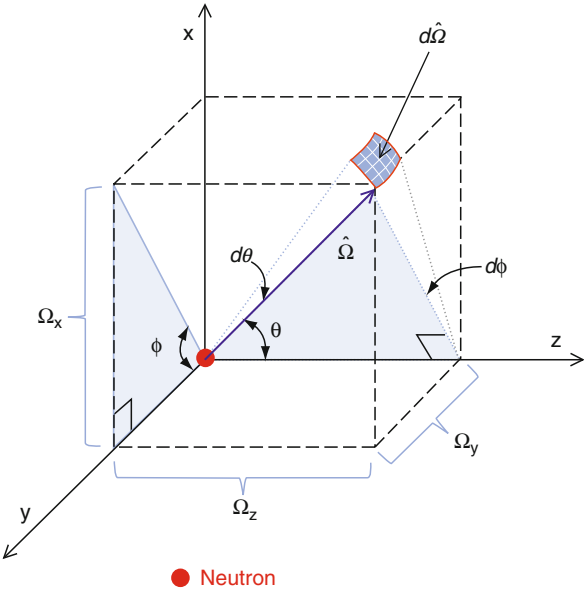
$$vN \equiv \Psi \text{particle flux, and} \quad (18)$$

$$\rho\sigma \equiv \Sigma \text{macroscopic cross section.} \quad (19)$$

The *flux* in our simple foil example is the number of particles per cm² per second crossing a plan parallel to the face of the foil. Given the more general representation of particle density (which will be used in the next section):

$N(\mathbf{r}, E, \mathbf{\Omega}, t) d\mathbf{r}^3 dE d\mathbf{\Omega} \equiv$ no. of particles in $d\mathbf{r}^3$ about $\mathbf{r}$, with kinetic energies in dE about E , and going in the solid angle $d\mathbf{\Omega}$ about the unit direction vector $\mathbf{\Omega}$ (see Fig. 5), at time t ; then the corresponding definition of *flux*, $\Psi(\mathbf{r}, E, \mathbf{\Omega}, t) ds dE d\mathbf{\Omega}$, is the no. of particles with E in dE going in direction $\mathbf{\Omega}$ in $d\mathbf{\Omega}$ that pass through the surface ds , which is located at $\mathbf{r}$ and is normal to $\mathbf{\Omega}$, per unit time, at time t .

The *macroscopic* cross section is the probability that a particle undergoes a reaction characterized by σ , per unit path (for small paths, dx) traversed by the particle in a homogenous material with target nucleus density, ρ . This definition lends a special significance to $1/\Sigma_T$, where $\Sigma_T = \rho\sigma_T$. σ_T is the *total* microscopic cross



Fission Reactor Physics. Figure 5 The unit direction vector $\mathbf{\Omega}$ associated with neutron velocity and the differential (small) solid angle $d\mathbf{\Omega}$ which defines the range of directions

section, the sum of the σ s for all of the reactions that the initiating particle can undergo with a given target isotope. So the change in flux, where $\mathbf{\Omega}$ is parallel to the x axis of a target material sample, over a small

interval dx in the sample is $d\Psi = -\Psi\Sigma_T dx$. And thus, $\Psi(x) = \Psi(0) e^{-\Sigma_T x}$, where x is the distance into the “sample” (which has its face in the $y-z$ plane at $x=0$). So, the probability of a reaction in dx about x can be expressed as

$$P(x)dx = \Sigma_T dx \bullet (\Psi(x)/\Psi(0)) = \Sigma_T e^{-\Sigma_T x} dx. \quad (20)$$

And thus the *mean free path* of a particle in an incident beam ($\Psi(0)$) before being removed from the beam in a homogeneous target is

$$\bar{x} = \int_0^{\infty} dx x P(x) = 1/\Sigma_T. \quad (21)$$

or generally the *mean free path* is the average distance traveled between successive interactions.

If a homogeneous material is made up of various nuclei (elements/isotopes, indexed by j), then the macroscopic cross section for a reaction, i , is

$$\Sigma_i = \sum_j \rho_j \bullet \sigma_i^j. \quad (22)$$

The reactions of primary interest in fission reactor design are those initiated by neutrons and gammas. Neutron cross sections are key to determining if a chain reaction can be maintained, and that the neutron population can be controlled under various transient conditions (e.g., start-ups and shutdowns, planned and accidental), and, of course, the fission distribution in the reactor. Most of the resulting energy release, from fission fragments, is deposited locally in the fuel elements of a given design. However, gammas, from fission and neutron capture in reactor structures, have large mean free paths, and their distribution and capture rates must be determined, using gamma cross sections, to complete the knowledge of energy deposition. The subsequent engineering problem is to assure that the reactor cooling system can remove the deposited energy under normal and accident conditions. Neutron and gamma cross sections are also required for the shield design of a fission reactor.

Neutron reactions are characterized by their energy balance, the Q factor, as well as microscopic cross sections. For the simple reaction (with the target at rest),



the energy balance is

$$(E_n + m_n c^2) + M_X c^2 = (E_Y + M_Y c^2) + (E_y + m_y c^2), \quad (24)$$

and Q is defined as the difference in the kinetic energies of the inputs (here the neutron) and the outputs:

$$Q = E_Y + E_y - E_n \text{ or } (M_X + m_n - M_Y - m_y)c^2. \quad (25)$$

If Q is positive, the reaction is exothermic, if negative, endothermic. For an endothermic reaction to go, for the microscopic cross section to be nonzero, enough kinetic energy must be supplied by the neutron to excite a compound nucleus, X^{A+1} , so it will decay to $Y + y$. As momentum must be conserved,

$$m_n v_n = (M_X + m_n) V_c \text{ or } V_c = v_n m_n / (M_X + m_n), \quad (26)$$

where V_c is the velocity of the compound nucleus. Then, the neutron energy supplied must be such that

$$-Q = m_n v_n^2 / 2 - (M_X + m_n) V_c^2 / 2 \quad (27)$$

and the *threshold energy*, E_{th} , for the reaction is

$$E_{th} = m_n v_n^2 / 2 = (-Q)(1 + m_n / M_X). \quad (28)$$

($n, 2n$) is an example of an endothermic reaction whose cross sections will exhibit an energy dependence of zero until the neutron energy E reaches an E_{th} .

The simplest, but very important, neutron reaction to be considered is a form of elastic scattering ($Q=0$), where collisions can be treated with classical mechanics as hard sphere, billiard ball, interactions. For the energies of neutrons in fission reactors, 0–10 MeV, elastic scattering cross sections for most nuclei are constant and proportional to the square of the nuclear radius, $\sim A^{2/3}$. Assuming the target nucleus to be at rest and applying conservation of energy and momentum in the center of mass, CM, coordinate system, one determines the probability that the final energy of the scattered neutron, in the laboratory coordinate, LM, system, is E_f in dE_f :

$$P(E_i \rightarrow E_f) = 1/(1-\alpha)E_i, \text{ for } \alpha E_i \leq E_f \leq E_i \\ = 0 \text{ otherwise,} \quad (29)$$

where $\alpha = ((A-1)/(A+1))^2$ and E_i is the initial neutron energy. A is the mass number of the target nucleus.

And, scattering is assumed to be isotropic in the CM coordinate system. This is a good assumption for the energy range of interest here, and its basis will be discussed later in this section. A full derivation of Eq. 29 can be found in Duderstadt and Hamilton, “Nuclear Reactor Analysis.” Examining $P(E_i \rightarrow E_f)$ one sees that a neutron scattering off a hydrogen nucleus ($A = 1$) can lose all its energy (as $\alpha = 0$). On average, it loses half its initial energy as

$$\bar{E}_f = \int_{\alpha E_i}^{E_i} dE_f E_f P(E_i \rightarrow E_f) = E_i(1 + \alpha)/2, \text{ and } (30)$$

$$\Delta \bar{E} = E_i \rightarrow \bar{E}_f = E_i(1 - \alpha)/2. (31)$$

Given $P(E_i \rightarrow E_f)$ as in Eq. 29, one defines *differential* microscopic elastic scattering cross sections, $\sigma_{es}^j(E_i)P(E_i \rightarrow E_f)dE_f$, which are particularly useful in determining how neutrons, born in fission, are slowed down in reactors designed to take advantage of the large fission cross sections of fissile isotopes in what is conventionally defined as the *thermal* neutron energy range, less than 0.625 eV. The superscript “j” of σ_{es}^j refers to the nuclei of the various *moderators* (hydrogen, deuterium and carbon) that are employed in these thermal reactors.

Once neutrons have slowed to the thermal range the target nuclei can no longer be assumed to be at rest. The interaction frequency will then be

$$|\mathbf{v} - \mathbf{V}| \sigma(|\mathbf{v} - \mathbf{V}|)\rho, (32)$$

where $|\mathbf{v} - \mathbf{V}|$ is the relative speed of neutron and target. For elastic scattering, $\sigma(|\mathbf{v} - \mathbf{V}|)$ is still nearly constant and an *average* cross section for *thermal* neutrons with speed v ($= (2E/m_n)^{1/2}$) is

$$\bar{\sigma}(v) = (\sigma_{es}/v\rho) \int d^3V |\mathbf{v} - \mathbf{V}| \rho(V), (33)$$

where for many reactor applications the Maxwell-Boltzmann velocity distribution for ideal gases in thermal equilibrium at absolute temperature, T , can be used to represent the targets. Thus,

$$\rho(V) = \rho \bullet (M/(2\pi kT))^{3/2} \exp(-MV^2/2kT), (34)$$

where M is the mass of the target nucleus and k is the Boltzmann constant (8.6174×10^{-5} eV/K, K is degrees Kelvin).

From Eq. 33, one sees that for $v \gg V$ the *average* cross section is, as expected, σ_{es} . And, as the neutron speed decreases and approaches zero, the *average* cross section goes as one over the neutron speed.

For highly accurate calculations (As part of the process of evaluating nuclear data sets, very accurate calculations of integral experiments are made. Zero power *mockups* of reactors, with carefully recorded dimensions and inventories, are commonly used. Monte Carlo calculations (to be discussed in the next section) of neutron balance in the mockups are made with various data sets (e.g., cross-section libraries) to determine a recommended set. See the CESWG web site for references to such experiments.), more sophisticated treatments of scattering from moderator structures (e.g., molecules in liquids, lattices for solids) are required. The excitation of modes of vibration, and thus energy loss to phonons must be considered. This has been a fertile field of development [9] and double differential scattering cross sections for various moderators have been produced. They are of the form:

$$\sigma_s(E_i \rightarrow E_f, \Omega_i \rightarrow \Omega_f) dE_f d\Omega_f = (1/4\pi kT) (E_f/E_i)^{1/2} \exp(-\beta/2) \sigma_{es} S(\alpha, \beta) dE_f d\Omega_f, (35)$$

where

$$\alpha \equiv (E_i + E_f - 2(E_i E_f)^{1/2}) \Omega_i \bullet \Omega_f / kT \text{ and } \beta \equiv (E_i + E_f) / kT. (36)$$

σ_{es} is the scattering cross section of the bound “moderating” nucleus (e.g., proton, deuteron, carbon). $S(\alpha, \beta)$, the *scattering law*, embodies the physics of the influence of the moderator structure on the scattering process. Various formulations of $S(\alpha, \beta)$ are tabulated as part of data files that document *all* the microscopic cross sections that are used in fission reactor design. These files can be found on the web site of the National Nuclear Data Center (currently, nndc.bnl.gov). The most widely used set is ENDF/B, the latest (2009) version is VII.0. In order, however, to produce the differential scattering cross sections (Eq. 35) for design calculations, material temperatures, T , must be identified and supplied with the corresponding ENDF files to NJOY [10], a system of computer programs which produce microscopic cross sections for use in various design programs (which will be discussed in the next section).

The remaining neutron reactions of interest all involve the formation of a compound nucleus, which will be in an excited state, $(X^{A+1})^*$, and will subsequently decay, yielding γ or γ 's (neutron capture), n (elastic neutron scattering), $n+\gamma$ (inelastic scattering), two n 's (an $(n,2n)$ reaction), p or α (charged particle production) or fissioning. The probabilities of these various outcomes for a given isotope, j , and incident neutron energy are characterized by the microscopic cross sections: $\sigma_c^j(E)$, $\sigma_s^j(E)$, $\sigma_{in}^j(E)$, $\sigma_{2n}^j(E)$, $\sigma_p^j(E)$, $\sigma_\alpha^j(E)$, and $\sigma_f^j(E)$. As noted above in the discussion of Q factors, for endothermic reactions, $Q < 0$, cross sections will be zero until a threshold value of E for the initiating neutron is reached. This is the case for inelastic scattering, $(n,2n)$ and some (n,α) and (n,p) reactions. There is similar threshold behavior for fertile isotope fission cross sections (see Sect. [Fission and Its Products](#)), which when the reactions “go” are exothermic. All neutron capture reactions (n,γ) are exothermic, and thus their cross sections are nonzero over the full range of fission reactor neutron energies.

One can view the “compound nucleus reaction” cross sections as the product of a cross section for compound nucleus formation, σ_C (neutron capture by the target nucleus), times the probability of a particular decay mode of the excited compound nucleus. Both factors of this “product” depend on the nature of the target and compound nuclei, X^A and X^{A+1} , and the energy available to excite the compound nucleus, X^{A+1} . The later is the sum of the reduced mass (i.e., center of mass) kinetic energy of the initiating neutron:

$$E_C = E(M_X/(m_n + M_X)) \cong E(A/(1 + A)), \quad (37)$$

where X^A is assumed to be at rest and momentum is conserved; and the excitation energy, E_e (see [Eq. 17](#)), provided by adding a neutron to X^A . E_e is the binding energy of the “added” neutron in the compound nucleus.

The magnitude of σ_C depends on the structure of X^A . First, if neutron number $N (=A-Z)$ is odd, σ_C is larger than its counterpart for neighboring isotopes with even neutron numbers. The opposite is true for N even. This just reflects the binding energy advantage of pairing half-integral spin Fermions in a nucleus (see the discussion of B_4 in Sect. [Fission and Its Products](#)). Second, for nuclei of various A 's there are *Magic*

Numbers for both Z and N (2, 8, 20, 50, 82 and 126) which can be thought of as closing shells of protons and neutrons, analogous to atomic electron shells. The reduction of σ_c for a magic number N nucleus, relative to its $N + 1$ isotope neighbor's σ_c , is much larger than the pairing effect.

Excited compound nuclei have mean lifetimes, τ (see [Eq. 14](#)) of as long as 10^{-14} s (Kaplan), much longer than the transit time for a neutron crossing a target nucleus, $\sim 2R/v$. Given the nuclear diameter, $2R \sim 10^{-12}$ cm, the transit time for even a *thermal* neutron is $\sim 10^{-17}$ s (The term *thermal neutron* refers to the most probable energy of the neutrons in thermal equilibrium in a zero-power reactor (e.g., a mockup). At 20°C, this is 0.023 eV, with a corresponding neutron speed of 2,200 M/s.). Thus, the standard assumption is that the decay of an excited compound nucleus is independent of all but the input energy of the initiating neutron. The decay modes of a particular excited state, nuclear level, are characterized by a *level width*

$$\Gamma \equiv h/(2\pi\tau), \quad (38)$$

with dimensions of energy (h is Planck's constant, 4.135667×10^{-15} eV-s), which is based on the Heisenberg uncertainty principle. In a quantum mechanical system like our excited compound nucleus, knowledge of energy and time is governed by

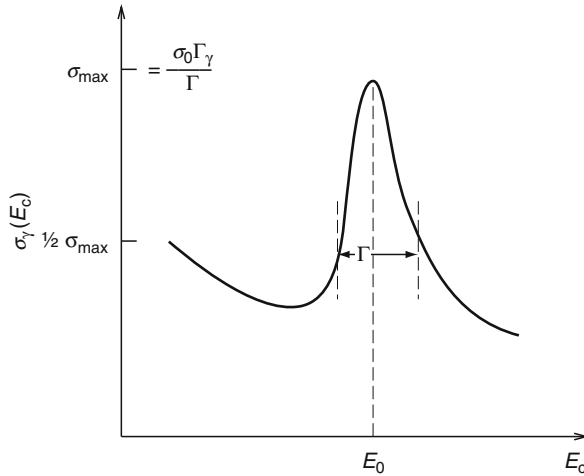
$$\Delta E \Delta t \sim h/2\pi. \quad (39)$$

Thus, Γ can be viewed as the uncertainty in energy of an excited state (level) of a compound nucleus, and τ a measure of the “uncertainty” of the lifetime of the excited state. The microscopic neutron cross sections, which go through the compound nucleus formation process, exhibit resonance behavior (peaking) when the neutron energy and E_e (the added neutron binding energy) produce or nearly produce a well-defined excited state (i.e., having a small Γ). See [Fig. 6](#).

The level width Γ can be thought of as the probability per unit time of decay of an excited state and thus the sum of partial “widths” (probabilities per unit time) for each mode of decay:

$$\Gamma = \Gamma_\gamma + \Gamma_n + \Gamma_{n\gamma} + \Gamma_f + \Gamma_{2n} + \Gamma_p + \dots, \quad (40)$$

(where Γ_n refers to elastic compound scattering and $\Gamma_{n\gamma}$ refers to inelastic scattering, the rest being obvious).



Fission Reactor Physics. Figure 6

A typical neutron capture cross section for an isolated (single) resonance whose width at half maximum is Γ , the total level width. Γ_γ is the partial width for gamma ray emission from the excited state (level). E_c is the center of mass (reduced mass) energy of the initiating neutron (Duderstadt)

Therefore, a “compound nucleus reaction” cross section near an isolated resonance is

$$\sigma(n, i) = \sigma_c(n) \Gamma_i / \Gamma, \text{ for } i = \gamma, n, (n\gamma), f, 2n, p, \text{ etc.} \quad (41)$$

The functional form of $\sigma(n, i)$, its dependence on neutron energy, was derived with the principles of quantum mechanics by Briet and Wigner [11] in 1935. Their “formula” for this simple case is

$$\sigma(n, i) = (\lambda^2 / 4\pi) \Gamma_n \Gamma_i / [(E - E_0)^2 + (\Gamma/2)^2], \quad (42)$$

where λ is the de Broglie wavelength of the neutron, $h/(2m_n E)^{1/2}$, and E_0 is the energy of the resonance peak. Figure 6 is an illustration of this “form” for $i = \gamma$.

Breit and Wigner’s most impressive derivation is more general than Eq. 42.

First, they considered neutron energies beyond what has been found pertinent to fission reactors. When one accounts for conservation of angular momentum, the initiating neutron has classically a magnitude of angular momentum $|L|$ equal to pb , where p is neutron momentum, $(2m_n E)^{1/2}$, and b is the

displacement of the neutron path from a parallel axis running through the center of the target nucleus. In a quantum mechanical treatment,

$$|L| = (l(l+1))^{1/2} h/2\pi \text{ where } l = 0, 1, 2, 3, \dots \quad (43)$$

Then, one can think of “ b ” as $|L|$ (given by Eq. 43) divided by the neutron momentum, p , and if there is going to be a significant probability of a reaction with the target nucleus, “ b ” cannot be much larger than the target nucleus radius, $r \cong A^{1/3}(1.28 \pm 0.05) \times 10^{-13}$ cm. For this to be true for a large nucleus, for example, for U^{235} , and for nonzero angular momentum (e.g., $l = 1$), the neutron would have to have kinetic energy ≥ 6.6 MeV. For smaller nuclei the required energy would be greater. Given the spectrum of neutrons in fission reactors, where most neutrons are born at around 2 MeV (see Fig. 4), an assumption of zero angular momentum ($l = 0$) for the vast majority of reactions is good, and thus equation Eq. 42 does not include a factor involving angular momentum or spin quantum numbers. This assumption also means that decay products of an excited compound nucleus will be released isotropically in the center-of-mass coordinate system, which is reflected in the factor of $1/4\pi$ in Eq. 42 (the probability that the decay product i ($i = n, \gamma, p$) is released $d\Omega$ about any Ω). The quantum mechanical treatment of angular momentum also accounts for the statement made above that “billiard ball” elastic scattering “can be assumed to be isotropic in the center of coordinate system.” This direct elastic scattering is referred to as *potential* scattering so as to be differentiated from resonance (compound nucleus) elastic scattering, that is, $i = n$ in equation Eq. 42.

Second, Breit and Wigner recognized and treated interference between potential and resonance elastic scattering. They found that the total elastic scattering cross section dips at energies right below the resonance peak, E_0 .

Finally, as they were aware that there could be multiple possible excited states of a compound nucleus they extended their “formula” to two resonances whose Γ s do not overlap.

Since Breit and Wigner’s original work, there has been great activity in measuring cross sections, motivated principally the desire to understand the physics

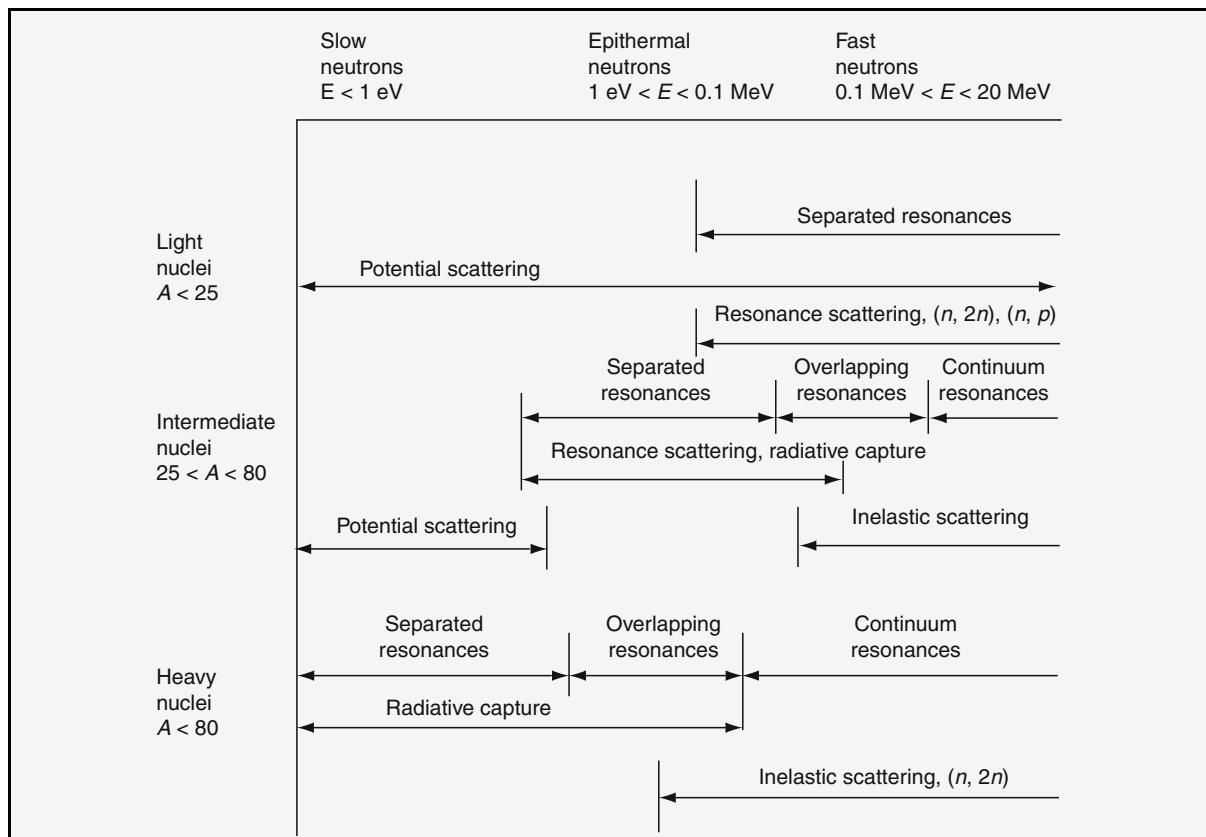
of the nucleus. In the process, however, the basic parameters required for nuclear weapon and reactor design were generated. The neutron cross sections for fission reactor design are summarized in Table 6 [12].

In this table, the distinction is made between resonance cross sections with different densities (spacing) of resonance peaks. With intermediate and heavier nuclei the level structure grows more complex, and the number of possible excited states of a compound nucleus greatly increases. With higher neutron energy more finely spaced excited states can be reached and their level width, Γ 's, increasingly overlap until measurement cannot resolve individual resonances.

In parallel with the work of nuclear spectroscopy experimentalists, theoreticians have built on Breit and Wigner's work. Resonance cross-section models [13] are key to creating Evaluated Nuclear Data Files.

The Cross Section Evaluation Working Group, a cooperative effort of national laboratories, industry and universities in the United States and Canada (see nndc.bnl.gov), sponsors reviews of the various measurements of a given cross section (target isotope and reaction) and the subsequent determination (As part of the process of evaluating nuclear data sets, very accurate calculations of integral experiments are made. Zero power *mockups* of reactors, with carefully recorded dimensions and inventories, are commonly used. Monte Carlo calculations (to be discussed in the next section) of neutron balance in the mockups are made with various data sets (e.g., cross-section libraries) to determine a recommended set. See the CESWG website for references to such experiments.) of a consensus set of parameters for an appropriate cross-section *model*. These models and their "consensus"

Fission Reactor Physics. Table 6 Types of neutron cross section for various target element/isotope masses pertinent to fission reactor design



parameters are a large part of the ENDF/B-VII.0 data files. Having the cross sections represented by an analytic model also facilitates dealing with the temperature effect on resonance cross-sections, that is, Doppler shift or broadening. The analytic process of averaging a resonance cross section (i.e., its model), over the velocity distribution of the target nuclei at a given temperature is similar to what was discussed above for reactions initiated by neutrons in thermal energy range. The process is outlined by Duderstadt and Hamilton using the single-level Breit Wigner formula as the resonance model. The effect of increasing temperature is to reduce a resonance peak while broadening its width, thus increasing its Γ . To first order, the area under the resonance is unchanged, which could lead one to think that resonance “Doppler” broadening is not an important effect in a reactor application. This is true if the density of the resonance target nuclei is small (i.e., it is very dilute in the reactor), and thus its presence does not change the energy dependence of the reactor’s neutron population. However, in most reactor designs, resonance absorbers are concentrated in localized reactor features (e.g., fuel elements, control rods) and there is significant self-shielding at the resonance peak. That is, neutrons with the “peak” energy will most likely be absorbed in the reactor “feature” irrespective of temperature-induced changes in the resonance microscopic cross section. But the story can be different on the wings of a resonance where the cross section is much smaller, and, thus, so is the self-shielding. An increase in temperature of the “feature” can result in a net increase in neutron absorption, with no change at the peak energy, but with increases in the wings. This phenomena can aid in insuring a *negative temperature coefficient* for a fission reactor design (*Temperature coefficients* are collective reactor parameters that reflect how neutron balance is impacted by temperature change through feedback mechanisms, for example, Doppler broadening (or narrowing) of resonances, and moderator density changes. Reactor control will be discussed in Sect. [Fission Reactor Performance](#)). A negative temperature coefficient is a crucial reactor design safety requirement.

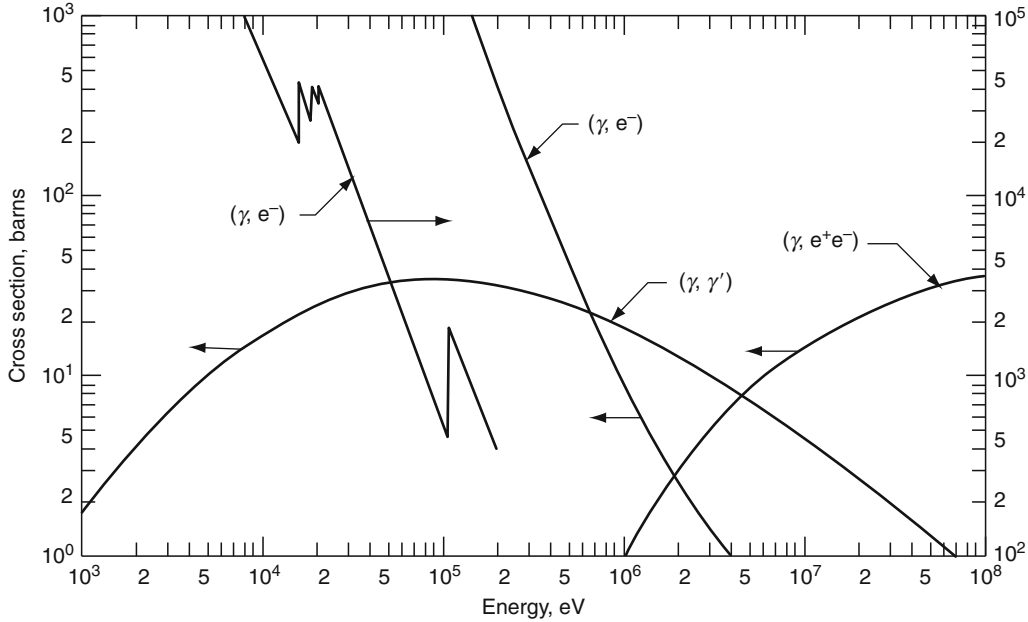
Dealing with temperature in producing resonance cross sections for design calculations is handled in a manner similar to the process for thermal energy

range cross sections discussed above. The ENDF resonance cross-section models and appropriate material temperatures are input to NJOY, and the output are the broadened cross sections in the model format. How these cross sections are used in design calculations is addressed in the next section.

As noted at the start of this section, gamma ray (photon) cross sections are important in fission reactor design as they are required for the full treatment of energy deposition throughout a reactor (in its fuel bearing *core* and supporting structures). In the order of importance with increasing gamma energy, the mechanisms of attenuation are the photo electric effect (γ, e^-), Compton scattering (γ, γ^*), and pair production (γ, e^-e^+). The photoelectric effect removes a γ when its energy, $h\nu$, can eject an atomic electron. Its cross section is approximately proportional to $Z^4/(h\nu)^3$, and has discontinuities in energy as the ionization energy of various atomic electron shells are achieved. For higher γ energies, Compton scattering interactions with atomic electrons can be treated as effectively free electron collisions. Conserving momentum and energy relativistically, one can derive expressions for energy loss and change in direction for initiating γ s as a function of their incident energy. The magnitude of the cross section is proportional to Z . When γ energies reach a threshold of 1.022 MeV ($2 \times m_e c^2$) and are in the field of a target nucleus, pair production of an electron and positron is possible. The magnitude of the pair production cross section is proportional to Z^2 . Of course, in tracking the γ population in a reactor, one recognizes that annihilation of a positron will produce two 0.51 MeV γ ’s. So pair production can be viewed as a form of inelastic scattering event. Cross sections for these three processes are tabulated in ENDF/B files, and they are described at a thorough but accessible level in the classic text by Robley D. Evans “The Atomic Nucleus.”

An example plot of these cross sections for Th^{232} is provided in [Fig. 7](#).

Finally, there are other gamma reactions which can take place in a reactor, for example (γ, f) and (γ, n) (the latter which we noted earlier as a source of delayed neutrons). However, these are threshold reactions for relatively high-energy gammas, and as shown in [Table 3](#) the total energy available from fission from fissile



Fission Reactor Physics. Figure 7

Thorium cross section for the photo electric effect (γ, e^-), for Compton scattering (γ, γ'), and pair production (γ, e^+e^-) [32]

isotopes for gammas is limited: <8 MeV for prompt γ 's, and, <6.5 MeV for delayed γ 's. Thus, these reactions are not important in determining the overall distribution of gammas in a reactor design.

Neutron Distributions

With the material provided to this point, the primary problem of reactor *theory* can be addressed: that of finding the neutron distribution in phase space ($\mathbf{r}, E, \Omega$), of a reactor design, and subsequently the reactor's power distribution, both throughout the reactor design's lifetime. The first task is to derive the equation for the neutron density, $N(\mathbf{r}, E, \Omega)$, the neutron transport equation, and auxiliary equations for the atom densities, $\rho_i(\mathbf{r}, t)$, of depleting (initial inventory) isotopes and important fission products (Like reactor physics, *reactor theory* is a traditional term in the nuclear engineering discipline. It refers to the study of the neutron transport equation and the means of its solution.). A simple approach to deriving the transport equation is to consider a balance relationship for $N(\mathbf{r}, E, \Omega, t)dEd\Omega dt$ for a time invariant volume, V , in configuration space:

$$\begin{aligned}
 dEd\Omega dt \int_V (\partial N(\mathbf{r}, E, \Omega, t)/\partial t) d^3r = \\
 & - \text{loss from flow out of } V \\
 & + \# \text{ scattering into } dE \text{ about } E \\
 & \text{and } d\Omega \text{ about } \Omega \\
 & + \# \text{ produced by fission} \\
 & + \# \text{ produced by other sources}
 \end{aligned} \quad (44)$$

For the "flow" term, one defines the neutron angular current $\mathbf{J}(\mathbf{r}, E, \Omega, t) \equiv \mathbf{v}N(\mathbf{r}, E, \Omega, t)$, where $|\mathbf{J}(\mathbf{r}, E, \Omega, t) \cdot \mathbf{n} ds dEd\Omega dt|$ is the no. of neutrons at $\mathbf{r}$, with energies in dE about E , traveling in $d\Omega$ about Ω , which cross an area ds with a unit normal vector $\mathbf{n}$ in dt at t . And thus, net flow out of V , which has a non-reentrant surface S , is:

$$\begin{aligned}
 dEd\Omega dt \int_S \mathbf{J}(\mathbf{r}, E, \Omega, t) \cdot \mathbf{n} ds \\
 = dEd\Omega dt \int_V \mathbf{v} \cdot \nabla N(\mathbf{r}, E, \Omega, t) d^3r,
 \end{aligned} \quad (45)$$

where Gauss' Theorem is applied to transform the surface integral to a volume integral.

The “reactions” in V are simply:

$$dE d\Omega dt \int_V \nu \Sigma_T(\mathbf{r}, E, t) N(\mathbf{r}, E, \Omega, t) d^3r, \quad (46)$$

where $\Sigma_T(\mathbf{r}, E, t)$ is the total *macroscopic* cross section in V . All scattering cross sections included in Σ_T have been integrated over final energies and directions.

Scattering into V is:

$$dE d\Omega dt \int_V d^3r \int_0^\infty dE' \int d\Omega' \nu' \Sigma_s(\mathbf{r}, t, E' \rightarrow E, \Omega' \rightarrow \Omega) N(\mathbf{r}, E', \Omega', t) \quad (47)$$

where Σ_s is the double differential *macroscopic* scattering cross section in V (see the definition Eq. 65 for an example of a double differential *microscopic* scattering cross section).

The direct fission source into V is:

$$dE d\Omega dt \int_V d^3r \int_0^\infty dE' \int d\Omega' \sum_i \nu_{pi}(E' \rightarrow E) \Sigma_{fi}(\mathbf{r}, E', t) \nu' N(\mathbf{r}, E', \Omega, t) / 4\pi, \quad (48)$$

where $\nu_{pi}(E' \rightarrow E)$ is the number of prompt neutrons emitted in dE about E by a fission of isotope i initiated by neutrons in dE' about E' : the macroscopic fission cross section for isotope “ i ” is $\rho_i(\mathbf{r}, t) \sigma_{fi}(E')$.

The delayed neutron source into V is:

$$dE d\Omega dt \sum_j X_{dj}(E) \frac{\lambda_j}{4\pi} \int_V C_j(\mathbf{r}, t) d^3r, \quad (49)$$

where $X_{dj}(E)/4\pi$ is the probability that the decay of delayed neutron precursor “ j ” will produce a neutron in dE about E and $d\Omega$ about Ω , and where λ_j and C_j are, respectively, the decay constant (see the beginning of section “[Heavy Elements](#)”) and isotope density of precursor “ j ”.

And, finally any source of neutrons in V not produced by a neutron reaction is given by:

$$dE d\Omega dt \int_V d^3r S(\mathbf{r}, E, t) / 4\pi, \quad (50)$$

where S could characterize a source (e.g., Plutonium (238)-Beryllium(α, n) or Californium-252 (spontaneous

fission)) included in a reactor design to aid in reactor start-up or S might account for decay processes yielding neutrons due to the presence of depleted fuel incorporated from another design.

Now, if the terms on the right side of the “balance relationship,” equation Eq. 44, are moved to the left side, and $dE d\Omega dt$ and the integral operation $\int_V d^3r$ is

factored out of all the terms, then as the right side is now zero and the small volume V is arbitrary, the collection of expressions under the integral must equal zero. The resulting equation is the *Neutron Transport (or Boltzmann) Equation*:

$$\begin{aligned} \frac{\partial}{\partial t} N(\mathbf{r}, E, \Omega, t) = & -\mathbf{v} \cdot \nabla N(\mathbf{r}, E, \Omega, t) \\ & -\mathbf{v} \Sigma_T(\mathbf{r}, E, t) N(\mathbf{r}, E, \Omega, t) \\ & + \int_0^\infty dE' \int d\Omega' \nu' \Sigma_s \\ & (\mathbf{r}, t, E' \rightarrow E, \Omega' \rightarrow \Omega) N(\mathbf{r}, E', \Omega', t) \\ & + \int_0^\infty dE' \int d\Omega' \nu' \sum_i \nu_{pi}(E' \rightarrow E) \\ & \Sigma_{fi}(\mathbf{r}, E', t) N(\mathbf{r}, E', \Omega', t) / 4\pi \\ & + \sum_j X_{dj}(E) \lambda_{dj} C_j(\mathbf{r}, t) / 4\pi \\ & + S(\mathbf{r}, E, \Omega, t). \end{aligned} \quad (51)$$

The conditions for solutions of this partial-differential-integral equation are the continuity condition:

$N(\mathbf{r} + \alpha \Omega, E, \Omega, t)$ must be a continuous function of α for $\mathbf{r} + \alpha \Omega$ in the reactor, and the

boundary condition:

$N(\mathbf{r}_s, E, \Omega, t) = 0$, for $\Omega \cdot \mathbf{n} < 0$, where $\mathbf{n}$ is an outward unit vector normal to a non-reentrant surface chosen to define the extent of the reactor.

The auxiliary equations for number densities of delayed neutron precursors, $C_j(\mathbf{r}, t)$, and fission product poisons, depleting fissile isotopes, and *burnable* poisons, $\rho_i(\mathbf{r}, t)$ s, are simply defined as movement of these isotopes in space can be ignored. *Burnable* poisons are elements with large neutron absorption cross sections (e.g., Boron, Hafnium, Cadmium,

Erbium, Gadolinium) that can be included in reactor designs to maintain neutron balance over design lifetime.

- For delayed neutron precursors:

$$\begin{aligned} \frac{\partial}{\partial t} C_i(\mathbf{r}, t) = & -\lambda_i C_i(\mathbf{r}, t) \\ & + \int_0^\infty dE' \int d\Omega' \sum_j \xi_{ij} \\ & v' \Sigma_{fj}(\mathbf{r}, E', t) N(\mathbf{r}, E', \Omega', t), \end{aligned} \quad (52)$$

where ξ_{ij} is the expected number of precursors of type i produced by fission of isotope j .

- For fission product poisons (e.g., Xenon-135 and its precursors Tellurium-135 and Iodine-135, and Samarium-149 and its precursors Neodimium-149 and Promethium-149):

$$\begin{aligned} \frac{\partial}{\partial t} \rho_i(\mathbf{r}, t) = & -\lambda_i \rho_i(\mathbf{r}, t) + \lambda_{j \rightarrow i} \rho_j(\mathbf{r}, t) \\ & - \int_0^\infty dE' \int d\Omega' \rho_i(\mathbf{r}, t) \sigma_{ci} \\ & (E') v' N(\mathbf{r}, E', \Omega', t) \\ & + \int_0^\infty dE' \int d\Omega' \sum_k \gamma_{ik}(E') \\ & \sum_{fk} (\mathbf{r}, E', t) v' N(\mathbf{r}, E', \Omega', t), \end{aligned} \quad (53)$$

where $\gamma_{ik}(E)$ is the expected number of poison (or poison precursor) nuclei produced by fission of isotope k induced by neutrons of energy E .

- For fissile fuel and burnable poison isotopes:

$$\begin{aligned} \frac{\partial}{\partial t} \rho_i(\mathbf{r}, t) = & - \int_0^\infty dE' \int d\Omega' \rho_i(\mathbf{r}, t) \sigma_{ai}(E') v' \\ & N(\mathbf{r}, E', \Omega', t), \end{aligned} \quad (54)$$

where the subscript “a” as applied to σ_a^i conventionally refers to capture plus fission for fissile isotopes. For reactor designs containing fertile isotopes, equation Eq. 54 will have a source term reflecting the transmutation process leading to the fissile isotope “i” (see Fig. 1). Additional “auxiliary”

equations may be needed to deal with transmutation of one isotope of a burnable poison to another, which has a significant neutron capture cross section (for example Hafnium, which has four naturally occurring isotopes).

Solving these “auxiliary” equations, irrespective of their number, is not a calculational challenge, given one knows the neutron density, $N(\mathbf{r}, E, \Omega, t)$, as they are ordinary differential equations. Obviously, solving the neutron transport equation (Eq. 51) for $N(\mathbf{r}, E, \Omega, t)$ is another matter. There are several features of fission reactors, however, that make this task more tractable. First, the density of neutrons needed to produce as much power as can be removed/transferred from various reactor types to do useful work is very small, $\sim 10^7 - 10^9 \text{ \#/cm}^3$, where as the density of nuclei is many orders of magnitude larger $\sim 10^{23} \text{ \#/cm}^3$. Therefore, the neutrons can be viewed as a very dilute gas in the matrix of a reactor’s nuclei, and thus neutron–neutron collisions can be ignored (they have not been accounted for in Eq. 51) (Another neutron behavior that can be ignored in formulating the transport equation is the finite lifetime of a free neutron. Its mean-life is $\sim 11.5 \text{ min}$. But, as will be discussed in the next section, the lifetime of neutron in a reactor is measured in milliseconds.). Second, in the primary nuclear design calculations, where it is determined if a trial configuration, loading and geometry, of fuel, structure, moderator, coolant, control elements, and poisons, can sustain neutron balance throughout the reactors lifetime objective, the time variable “t” in Eq. 51 can be treated in a much simplified manner. Given the initial conditions of a trial reactor configuration, it is a good assumption that the atom densities in the various macroscopic cross sections in Eq. 51 can be treated as nearly constant for a significant time interval, $\Delta t \geq \text{minutes}$. With this assumption, and no neutron–neutron collisions, Eq. 51 is linear in $N(\mathbf{r}, E, \Omega, t)$ during Δt , and solution methods for Eq. 51 are greatly simplified. In addition, for the interval, Δt , Eq. 51 can be treated as a time-independent equation. For a *primary nuclear design* calculation, one ignores the source term $S(\mathbf{r}, E, t)$ (its importance to the start-up problem will be addressed in the next section) and Eq. 51 becomes a linear homogeneous eigenvalue problem:

$$\begin{aligned}
& \mathbf{v}\Omega \bullet \nabla \cdot \mathbf{N}(\mathbf{r}, E, \Omega) + \nu \Sigma_T(\mathbf{r}, E) \mathbf{N}(\mathbf{r}, E, \Omega) \\
& - \nu \int_0^\infty dE' \int d\Omega' \nu' \Sigma_s(\mathbf{r}, E' \rightarrow E, \Omega' \rightarrow \Omega) \mathbf{N}(\mathbf{r}, E', \Omega') \\
& = \frac{1}{k} \int_0^\infty dE' \int d\Omega' \nu' \sum_j \Sigma_{fj}(\mathbf{r}, E') \\
& \left\{ \nu_{pj}(E' \rightarrow E) + \sum_j X_{di}(E) \xi_{ij} \right\} \mathbf{N}(\mathbf{r}, E', \Omega') / 4\pi,
\end{aligned} \tag{55}$$

where $1/k$ is the eigenvalue. It has been customary to use the inverse of “ k ” as the eigenvalue and to refer to k as the multiplication factor. Note that if Eq. 55 is integrated over the reactor volume and E and Ω , then k is equal to the ratio of neutron production to neutron loss, this is the origin of its designation as a “multiplication factor.” A further simplification in Eq. 55 arising from the assumption of “time independence” during Δt , is that delayed neutron production can just be added to prompt neutron production. As indicated, previously delayed neutrons, though less than a percent of total neutron yield in a fission, are critical to transient reactor behavior, and, therefore, control system design, to be covered in the next section.

Before proceeding with the solution methods for Eq. 55, a description of how the *primary nuclear design* calculation proceeds is required for a basic understanding of reactor design, and to provide perspective on the utility of the various solution methods for the transport equation. Assuming the solution method chosen has yielded a k_i and $\mathbf{N}_i(\mathbf{r}, E, \Omega)$ for the initial time interval, Δt_i where $i = 1$, then one proceeds to check design requirements, and make needed modifications to the reactor’s initial *trial configuration*. In general, this is an iterative process including other disciplines (i.e., heat transfer and fluid flow, structural analysis). If $k \neq 1$, inventories of fuel/poisons or the positioning of control elements will need to be altered to achieve neutron balance (*Control elements* (sometimes generically called control rods) are structures incorporating highly neutron absorbing isotopes. They have a dual role in fission reactor design: (1) assuring criticality (a controlled steady state chain reaction) throughout a reactor’s design lifetime, i.e., compensating for potential excess neutron production from the initial fissile loading which must be large enough to accommodate

depletion; and (2) providing safe shut down (termination of the chain reaction) of the reactor in case of an accident condition or during a planned interruption of normal operations (e.g., for plant maintenance or refueling). The same “structures” could accomplish both functions or there could be separate structures (sometimes referred to as “shim” and “shutdown” rods respectively).). The power distribution throughout the reactor must then be determined. As $\mathbf{N}(\mathbf{r}, E, \Omega)$ is the solution to a homogeneous equation, its absolute magnitude is undetermined, but a normalization factor, p , can be established from the total thermal power rating, $P(\text{MWth})$, requirement of the design, that is from:

$$\mathbf{P} = p \int_{\text{RVol.}} d^3r \int_0^\infty dE \int d\Omega \nu \sum_j \varepsilon_j \Sigma_{fj}(\mathbf{r}, E) \mathbf{N}(\mathbf{r}, E, \Omega), \tag{56}$$

where ε_j is the energy release per fission of isotope j (see Table 3). Now given $p\mathbf{N}(\mathbf{r}, E, \Omega)$, one can calculate the power distribution throughout the reactor

$$\mathbf{P}(\mathbf{r}) = p \int_0^\infty dE \int d\Omega \nu \sum_j \varepsilon_j \Sigma_{fj}(\mathbf{r}, E) \mathbf{N}(\mathbf{r}, E, \Omega), \tag{57}$$

and determine peak powers in fuel elements, and average and peak heat fluxes into coolant channels. Thus, fuel element and heat removal system limits can be checked. If there are violations, the *trial configuration* must be altered and new results for k and $\mathbf{N}(\mathbf{r}, E, \Omega)$ found. From the power distribution and subsequent thermal analysis, one can also verify the temperatures that were assumed for the *trial configuration*, that is, the temperatures that were needed to define thermal scattering cross sections and Doppler broadened resonance cross sections. Finally, when all conditions are met for the first time interval, Δt_i , where $i = 1$, the *auxiliary* equations (Eq. 53 and Eq. 54) can be solved to update inventories of fuel and poisons for the next time interval, Δt_2 . For a thermal reactor design, where the fission product poisons Samarium-149 and Xenon-135, are important, initial time intervals should be short (minutes) until equilibrium levels of these isotopes are achieved ($\sim$ hours). Subsequent time intervals can be many hours. From this brief description of the *primary nuclear design process*, it should be clear that

having accurate and efficient solution methods for the time-independent neutron transport equation is key to achieving a successful design.

There are two basic approaches to solving the time independent neutron transport equation, Eq. 55; the probabilistic-statistical Monte Carlo simulation method [14] where individual neutrons are traced through their life experience in a reactor, and analytic methods where the variables, $\mathbf{r}$, E , and (sometimes) Ω , are made discrete, thus transforming Eq. 55 into a matrix equation.

Monte Carlo simulation is in principle well suited for this application because the neutron density is low so the transport equation can be treated as linear. Thus, each “experiment,” that is, neutron history, is independent of all others. Initially, a neutron is started at a randomly selected reactor location and with a randomly selected direction, Ω . Its initial energy is selected by treating a typical prompt neutron energy spectrum as a probability distribution:

$$P(E)dE = v_p(E)dE / \int_{E_{\min}}^{E_{\max}} dE v_p(E), \quad (58)$$

which is then “sampled” with a random number between 0 and 1. In Monte Carlo computer programs [15, 16] a sequence of random numbers is generated with an algorithm. To provide an example of “sampling,” note that in this case, given a random number n , the “sampled” starting energy E is found by simply solving the transcendental equation:

$$n = \int_{E_{\min}}^E dE' P(E') \text{ for } E. \quad (59)$$

Now having a speed and direction, one can “sample” (with a new random number) the probability that the neutron travels a distance x before having a collision, using equation Eq. 20. As this probability depends on the total macroscopic cross section, $\Sigma_T(\mathbf{r}, E)$, one must keep track of material boundaries. An initial sampling will be for the distance from the starting point of the neutron to the first material boundary it could cross. If the initial sampling results in the boundary being crossed, then there is a new Σ_T and a second sampling is performed to determine if

another boundary is crossed. Eventually, either the neutron leaves the reactor or the location of the first collision is established. As the total cross section is the sum of capture, scattering, and possibly fission macroscopic cross sections for the various isotopes present, one can treat the relative magnitudes of the components of the sum as a probability distribution, which when “sampled” leads to the next step in our neutron’s history. If capture is the result, the history ends, just as it would end if the neutron leaked (escaped) from the reactor. Either capture or leakage is recorded as a “loss.” If fission is the result, the history also ends, but the number of neutrons produced (1, 2, or 3) in the fission of the “selected” isotope j , and the fission location are recorded. The number of fission neutrons is determined by “sampling” the probability distribution for fission yield of isotope j . The mean of the distribution $\bar{\nu}_j(E)$ includes delayed neutrons. The number of neutrons produced by a “history” is counted as “production.” If scattering is the result of the collision then the double differential macroscopic cross section is treated as probability distribution and “sampled” to provide a new energy and direction for the neutron. Then, the process of tracking to either escape or the next collision is repeated. Thus, the “history” proceeds until it ends as either “loss” or “production”, if “production” a starting point for a subsequent “history” and a location for energy deposition are also provided.

There are various strategies for carrying out Monte Carlo calculations and the evaluation of the statistical nature of the results. A standard approach, in outline, is to run successive groups of “histories” (thousands). Discard the first few groups, which are used to establish a reasonable fission neutron spatial source distribution, and then find the mean and standard deviations of the desired calculational results from the subsequent groups. It is most economical to get good statistics for the eigenvalue k , the multiplication factor, which is just “production” over “loss,” quantities which are accumulated over the whole reactor. Energy deposition, that is, power distribution, results are much more costly. Hundreds of groups with millions of histories per group would be required to give good statistics (a five percent 95% confidence interval) for the number of fissions in small reactor volumes (e.g., a 1 cm length of a typical PWR fuel element which has a volume of 0.7 cm^3 , out of a reactor volume of $32.8 \times 10^6 \text{ cm}^3$

(for a 3,400 MWth rating)). From this discussion, one can see why, as mentioned in section [Cross Sections](#), Monte Carlo calculations of mock-up experiments are widely used in cross-section data set evaluations, where the results of interest are changes in k , the multiplication factor. Even with the tremendous advances in computing capability which have been made to date, Monte Carlo simulation is not as yet the main line method for primary nuclear design calculations. But, as will be seen in the following description of analytic methods, it can greatly aid in improving the accuracy of the analytic methods.

When the analytic approach of making the neutron density's variables $\mathbf{r}$, E and Ω discrete is applied, so as to make the computational errors in solving the transport equation comparable to an exhaustive Monte Carlo simulation, the computer resource requirements will challenge today's largest machines (peak speeds of $\sim 2.3 \times 10^{15}$ flops (floating point operations per second) [17]).

This can be demonstrated with the large PWR used in the Monte Carlo discussion: *First*, a spatial mesh of 65.6×10^6 points would result, assuming quarter core radial symmetry, and from using a $1 \times 1 \times 2 \text{ cm}^3$ cell for averaging cross sections. (The mesh may need to be finer for highly absorbing features, e.g., fixed poisons, control elements, and can be coarser in homogeneous regions.) *Second*, the energy variable can be treated with a multigroup approximation where the energy range, $0.0 \rightarrow 10 \text{ MeV}$ is divided into intervals (groups), for example, 24 for thermal neutrons, $0.0 \rightarrow 0.625 \text{ eV}$, and 57 for the rest of the range, with most of these groups allotted to epithermal neutrons, $E = 0.625 \text{ eV} \rightarrow 0.1 \text{ MeV}$ where there is a concentration of explicit resonance cross sections (see [Table 6](#)). A weight function $f_g(\mathbf{r}, E)$, normalized for E to unity, is required for each interval, g . These can be calculated with infinite medium problems representing various portions of the reactor; homogeneous problems with no spatial variables, or an infinite repeating array of cells (e.g., a fuel element and surrounding coolant) with radial spatial variables, but no axial, z , dependence. The power of the multigroup approximation is that it is insensitive to the choice of weight functions and thus simplifying assumptions can be made in selecting and solving the "infinite medium" problems to generate the f_g 's.

Finally, the direction variable, Ω , can be dealt with through a discrete ordinate approximation [18]. The unit sphere is divided into segments, the surface areas of which sum to 4π , and a direction vector Ω_n is assigned to each segment. The segment surface areas act as weight functions when the transport equation is integrated over a segment to yield an equation for the neutron density going in the direction Ω_n . There are various schemes for selecting ordinates and weights, the most widely known is the S_n method. All methods, however, can produce *ray effects* if "n" is too small [19]. The channeling of neutrons into a few restricted directions can produce anomalous results in reactor designs with localized neutron source regions, that is, where fissile and fertile fuel predominate in different regions (commonly referred to as seed-blanket designs). For a "highly accurate" treatment, one should let $n = 16$ in each octant, for a total of 128 ordinates. So for the "large PWR example," the number of unknowns to be solved for in the discretized time-independent transport equation is $170 \times 10^9 (= 16.4 \times 10^6 \text{ (spatial mesh points)} \times 81 \text{ (energy groups)} \times 128 \text{ (ordinates)})$. This is clearly a formidable calculational problem. If we view the analytic approach described here as transforming [Eq. 55](#) into a matrix eigenvalue equation, then the simplest solution method of matrix inversion would involve matrices of a billion by a billion. Hence, an iterative method is required [20]. Much effort in reactor *theory* has been devoted to this problem, and to simplifying the analytic approach. Iterative methods for solving matrix equations are beyond the scope of this entry, but to understand how fission reactor design is actually carried out, a description of analytic approach simplifications is needed.

The direction variable, Ω , received the earliest attention. Because the neutron population in a reactor can be viewed as a dilute gas, it was natural to assume that the variation of the neutron density in space could be approximated by $\nabla \bullet D(\mathbf{r}, E, t) \nabla v N(\mathbf{r}, E, t)$ (from Fick's Law of diffusion). When the transport equation, [Eq. 51](#), is integrated over Ω and the first term on the right-hand side is replaced by the Fick's Law expression, the result is the time-dependent neutron diffusion equation. [Equation 55](#) can be treated analogously to yield the time-independent neutron diffusion equation. In either case, the limitations of the diffusion approximation

only become apparent in trying to define the diffusion coefficient $D(\mathbf{r}, E, t)$ (see Henry, Nuclear-Reactor Analysis). The most commonly used expression is

$$D(\mathbf{r}, E, t) \equiv 1/\{3[\Sigma_t(\mathbf{r}, E, t) - \bar{\mu}_0 \Sigma_s(\mathbf{r}, E, t)]\}, \quad (60)$$

where $\bar{\mu}_0$ is the average of the average cosine of the scattering angle (in the laboratory coordinate system) of each isotope making up $\Sigma_s(\mathbf{r}, E, t)$. This definition (Eq. 60) arises from a low-order spherical harmonics expansion of $N(\mathbf{r}, E, \Omega, t)$ in the neutron transport equation. Spherical harmonics are a complete orthonormal set of special functions on the unit sphere defined by Ω (the unit vector which is at angle θ from the z axis and projects in the x - y plane at angle φ from the x axis);

$$Y_l^m(\Omega) = H_l^m P_l^m(\mu) e^{im\varphi} \text{ for } l = 0, 1, 2, 3, \dots \\ m = -l \leq m \leq l, \quad (61)$$

where

$H_l^m = \left[\frac{(2l+1)(l-m)!}{4\pi(l+m)!dm} \right]^{1/2}$, $P_l^m(\mu) = \sin^m(\mu) \frac{d^m}{d\mu^m} P_l(\mu)$ is the associated Legendre Polynomial, $P_l(\mu) = \frac{1}{2^l l!} \frac{d^l}{d\mu^l} (\mu^2 - 1)^l$ is the Legendre Polynomial and $\mu = \cos\theta$. The spherical harmonics are normalized by the relationship:

$$\int_0^{2\pi} d\varphi \int_{-1}^1 d\mu \overline{Y_l^m(\mu, \varphi)} Y_{l'}^{m'}(\mu, \varphi) = \delta_{mm'} \delta_{ll'}, \quad (62)$$

where $\overline{Y_l^m}$ is the complex conjugate of Y_l^m and the Kronecker deltas, δ , are 1 for $m = m'$ and $l = l'$ and 0 otherwise.

The “low-order” spherical harmonics expansion, which yields the diffusion approximation, is for $l = 0$ and 1 (also referred to as the P1 approximation). The four functions in the expansion are:

$$Y_0^0 = (1/4\pi)^{1/2}, Y_1^{-1} = (3/4\pi)^{1/2} \sin \theta e^{-i\varphi}, \\ Y_1^0 = (3/4\pi)^{1/2} \mu \text{ and } Y_1^1 = -(3/4\pi)^{1/2} \sin \theta e^{i\varphi}. \quad (63)$$

In a Cartesian coordinate system, the expansion coefficients (for simplicity of notation the time-independent case will be treated) are $N(x, y, z, E)$, the

neutron density and $N_1^{-1}(x, y, z, E)$, $N_1^1(x, y, z, E)$ and $N_1^0(x, y, z, E)$, which when multiplied by v (the speed corresponding to E) are the neutron currents in the x , y , and z directions. Before applying the P1 expansion to Eq. 55, one needs to note that the double differential scattering cross section, $\Sigma_s(E' \rightarrow E, \Omega' \rightarrow \Omega)$ in reactor applications (where neutron polarization and Bragg scattering can be ignored) depends on $\Omega' \bullet \Omega$. Then, given the addition theorem for spherical harmonics,

$$P_1(\Omega' \bullet \Omega) = \sum_{m=-1}^1 \frac{(l-m)!}{(l+m)!} P_l^m(\mu) P_l^m(\mu') e^{im(\varphi-\varphi')}, \quad (64)$$

$\Sigma_s(\mathbf{r}, E' \rightarrow E, \Omega' \bullet \Omega)$ can be expanded in terms of associated Legendre polynomials. This is done in two steps: first the double differential scattering cross section is expressed as an expansion in ordinary Legendre polynomials (which are a complete orthogonal set of functions)

$$\Sigma_s(\mathbf{r}, E' \rightarrow E, \Omega' \bullet \Omega) = \Sigma_s(\mathbf{r}, E) \sum_{l=0}^{\infty} \frac{(2l+1)}{4\pi} F_l(\mathbf{r}, E' \rightarrow E) P_l(\Omega' \bullet \Omega), \quad (65)$$

and then Eq. 64 is substituted for $P_l(\Omega' \bullet \Omega)$. (It should be remembered that Σ_s is a macroscopic cross section, and a more complete notation would show a sum over contributions from each isotope present in d^3r about $\mathbf{r}$.)

Now when the P1 expansion for $N(x, y, z, E, \mu, \varphi)$ is inserted in Eq. 55 and the resulting equation is, in turn, multiplied by the complex conjugate of each of the four spherical harmonics functions of the P1 expansion (Eq. 63), and integrated over μ ($-1 \rightarrow 1$) and φ ($0 \rightarrow 2\pi$), four equations result: The first is

$$\frac{\partial}{\partial x} v N_1^{-1}(x, y, z, E) + \frac{\partial}{\partial y} v N_1^1(x, y, z, E) \\ + \frac{\partial}{\partial z} v N_1^0(x, y, z, E) = -v \Sigma_T(x, y, z, E) N(x, y, z, E) \\ + \int_0^{10\text{Mev}} dE' v' \Sigma_s(x, y, z, E') F_0(E' \rightarrow E) N(x, y, z, E') \\ + (\text{the fission term in Eq.55 with } N(\mathbf{r}, E', \Omega') \\ \text{replaced with } N(x, y, z, E')). \quad (66)$$

The remaining three equations are of the same form, one is provided here:

$$\begin{aligned} \frac{1}{3} \mathbf{v} \frac{\partial}{\partial z} N(x, y, x, E) + \nu \Sigma_T(x, y, z, E) N_1^0(x, y, z, E) \\ = \int_0^{10 \text{ MeV}} dE' \nu \Sigma_s(x, y, z, E') F_1 \\ (E' \rightarrow E) N_1^0(x, y, z, E') \end{aligned} \quad (67)$$

With these four coupled partial-differential-integral equations, there is still considerable computational complexity. To get to the standard neutron diffusion equation, one additional approximation is made:

$$F_1(E' \rightarrow E) \cong \delta(E' - E) \mu(E'), \quad (68)$$

where $\mu(E')$ is the average cosine of the scattering angle in the laboratory coordinate system, and energy loss (or gain) in the non-isotropic component of scattering is ignored. Substituting Eq. 68 in Eq. 67, the relationship between N_1^0 and N is a simple partial differential equation:

$$N_1^0(x, y, z, E) = - \frac{1}{3(\Sigma_T(x, y, x, E) - \mu(E) \Sigma_s(x, y, z, E))} \frac{\partial}{\partial z} N(x, y, z, E), \quad (69)$$

containing the diffusion coefficient $D(\mathbf{r}, E)$ (see the definition Eq. 60). Given the equations of the same form as Eq. 69 for N_1^1 and N_1^{-1} , and substituting all three into the left-hand side of Eq. 66. The left-hand side becomes the standard Fick's Law expression:

$$-\nabla \bullet D(\mathbf{r}, E) \nabla N(\mathbf{r}, E) = \dots \quad (70)$$

As noted above, this derivation of the diffusion approximation reveals its limitations. The ability of some combination of four low-order spherical harmonic functions, Eq. 63, to describe the true angular distribution of the neutron density throughout a reactor will be limited to regions where the distribution is nearly isotropic, that is away from boundaries and highly absorbing features (control elements and fixed poisons). To address these limitations, special boundary conditions are used, and subsidiary

calculations (to be discussed below) are made to provide “fitted” cross sections for highly absorbing features.

To make the diffusion approximation an efficient design tool, additional simplifications have been developed. The differencing of the energy variable as described above for “multigroups” can be extended to a “few group” approximation. Again, weight functions are generated using accurate solutions to small region “cell” problems, which model repeating features of a reactor. But here the weight functions are applied over much larger energy ranges, three or four to cover the energy range of 0 to 10 MeV. Furthermore, one accepts the error associated with the weight functions not perfectly representing the spatial variation of neutron density energy dependence.

The use of “cell problem” auxiliary calculations can be extended. As the core, the central fuel bearing region, of most reactor designs is made up of collections of mostly fuel elements, and possibly some fixed poison and movable control elements, assembled into modules. One can perform highly accurate (e.g., Monte Carlo or multigroup, fine spatial mesh, discrete ordinate) calculations for two-dimensional (radial) repeating arrays representations of a core's various modules. Then, for each module type, a series of corresponding few-group diffusion approximation calculations can be carried out, in which key few group cross sections are adjusted (“fitted”) to match reaction rates from the “highly accurate” reference calculation. One can use these fitted cross sections in a full core three-dimensional few group diffusion calculation as part of the *principle design process*.

Of course, with depletion, as inventories of fuel, fission products, and poisons are updated, fitting calculations will have to be repeated. This approach is particularly suited to the design of thermal reactors, that is, PWRs and BWRs, where the proper treatment of epi-thermal and thermal neutrons, including the effects of self-shielding by fuel and poison inventories, is crucial.

In fast reactor design, where most fissions take place at energies above explicit resonances (liquid metals or gasses are coolants, and fissile and fertile densities and inventories are high (no effective moderators are present)) and the mean free path of fission neutrons is large (~ 10 cm.), the treatment of fuel, structure and

coolant as an homogenized material, in formulating macroscopic cross sections, is a reasonable assumption. Treatment of the energy variable is also simpler as up-scattering is not important. As a result, multigroup, 3D discrete ordinate calculations can be used in the *principle design process* for fast reactors.

Finally, returning to thermal reactor design, there is another analytic calculational approach, which is in wide use, that should be mentioned. The general designation is Nodal Methods. There are a number of variations under this title [21], but they all have as their starting point solving 2D module array problems. One needs to produce few group neutron distributions within each module. Each module (or depending on symmetry each half or quarter module) will be a “node.” Then coupling coefficients between nodes, both radially and axially, are generated. (It is predominately in defining coupling coefficients that the various “methods” differ.) The resulting nodal equations can be solved with modest computer resources, but, to obtain power distributions and to update inventories the full reactor neutron distribution must be constructed from the module solutions and the weights found for each node. Nodal Methods have been found to be particularly useful in applying the *primary nuclear design process* to evaluating refueling options for commercial (large scale) PWRs and BWRs, where partially depleted modules are relocated, “shuffled,” as new modules are added and fully depleted, “spent” modules removed during periodic refuelings. The computational economy of nodal methods also allows them to be applied to fully time-dependent problems, particularly for accident analyses. The nature of these problems will be discussed in the next section.

Fission Reactor Performance

In the previous section, the focus was the derivation of the neutron transport equation, and how it is solved in carrying out the *primary nuclear design process*. This quasi-static *process* involves a series of time-independent calculations of the neutron density, $N(\mathbf{r}, E, \boldsymbol{\Omega})$, and ultimately results in the configuration and inventories (loadings) that meet design requirements for lifetime (total energy production), and normal operation thermal performance (fuel element burn-up within limits and sufficient coolant flow

provided by the design pumping power allocation). There are, however, additional design requirements that involve transients, that is, $N(\mathbf{r}, E, \boldsymbol{\Omega}, t)$, which will be the subject here.

The simplest approach to treating transient reactor behavior is through a “point” kinetics model. If one first multiplies the time-dependent transport equation, Eq. 51, by a weight function, $W(\mathbf{r}, E, \boldsymbol{\Omega})$, and integrates over the reactor volume, energy (the full range, $0 \rightarrow 10$ MeV), and direction ($\cos \theta$ from -1 to 1 , φ from 0 to 2π); and second, multiplies the time-dependent equation for each delayed neutron precursor, Eq. 52 for $C_i(\mathbf{r}, t)$, by $W(\mathbf{r}, E, \boldsymbol{\Omega})X_{di}(E)$ and performs the same integration over reactor volume, energy, and direction, the results are the point kinetics equations:

$$\frac{dT(t)}{dt} - \frac{\rho - \beta}{\Lambda} T(t) + \sum_{i=1}^I \lambda_i C_i(t) + Q(t) \quad (71)$$

$$\frac{dC_j(t)}{dt} = \frac{\beta_j}{\Lambda} T(t) - \lambda_j C_j(t) \text{ for } j = 1, 2, \dots, I \quad (72)$$

where $T(t)$ is an amplitude function;

$$T(t) \equiv \int_{\text{reactor}} dV \int_0^{10\text{Mev}} dE \int d\Omega W(\mathbf{r}, E, \boldsymbol{\Omega}) N(\mathbf{r}, E, \boldsymbol{\Omega}, t). \quad (73)$$

In formulating the expressions for the kinetics parameters, $\rho(t)$, $\beta(t)$, and $\Lambda(t)$, it is convenient to factor the neutron density into a product of “shape” and amplitude functions. The shape function is;

$$S(\mathbf{r}, E, \boldsymbol{\Omega}, t) \equiv N(\mathbf{r}, E, \boldsymbol{\Omega}, t)/T(t). \quad (74)$$

Now the weight function is defined over the same domain (space, energy, and direction) as the neutron density, and thus from the definitions Eq. 73 and Eq. 74 the normalization of S and W follows:

$$\int_{\text{reactor}} dV \int_0^{10\text{Mev}} dE \int d\Omega W(\mathbf{r}, E, \boldsymbol{\Omega}) S(\mathbf{r}, E, \boldsymbol{\Omega}, t) = 1 \text{ for all } t. \quad (75)$$

In order for the point kinetics equations to provide accurate solutions for small changes in reactor configuration, the weight function, $W(\mathbf{r}, E, \boldsymbol{\Omega})$, is chosen to be the solution the adjoint equation corresponding to the time-independent transport equation, Eq. 55, for the

reactor of interest adjusted to be critical (i.e., the eigenvalue $k = 1$). In the adjoint of Eq. 55, the variable pairs (E, Ω) and (E', Ω') are interchanged in the scattering and fission terms. The solution, $N^*(\mathbf{r}, E, \Omega)$, is referred to as the adjoint neutron density or the importance function. The latter name indicates the physical interpretation of N^* . If the reactor described by Eq. 55 is at zero power (no neutrons) and a neutron is inserted at $\mathbf{r}$ with velocity $\mathbf{v}(E, \Omega)$, the neutron level will increase to a steady-state value (remember the reactor is still critical). This “level” per neutron added at $(\mathbf{r}, E, \Omega)$ is $N^*(\mathbf{r}, E, \Omega)$. How N^* , acting as a weight function, improves the point kinetics equations will be discussed after ρ , β , and Λ are defined and their physical interpretation given. To simplify notation, the scattering and fission integral operators in Eq. 51 are represented by A and G :

$$A \equiv \int_0^\infty dE' \int d\Omega' v' \Sigma_s(\mathbf{r}, E' \rightarrow E, \Omega' \rightarrow \Omega) \bullet, \quad (76)$$

$$G \equiv \int_0^\infty dE' \int d\Omega' v' \sum_i v_i(E') \Sigma_{fi}(\mathbf{r}, E') \bullet, \quad (77)$$

where $v_i(E')$ is the total number of neutrons (prompt plus delayed) produced by fission of isotope i induced by a neutron with energy E' . In reactor kinetics, delayed neutron yields are expressed in terms of the ratio the number produced with a given half-life (i.e., a member of “delayed group,” j , as noted previously, see Table 4) by fission of isotope, i , to the total the total yield, v_i . These ratios are represented as β_j^i , where normally $j = 1, 2 \dots 6$.

The parameter $\rho(t)$ is the *reactivity* of a reactor and is a measure of how far from criticality (a steady-state chain reaction only from fission neutrons, no other neutron sources present) the reactor is at time t . This can be seen from the expression for $\rho(t)$ that results from the derivation of Eq. 71 from the transport equation (where the functional dependencies on $\mathbf{r}$, E , Ω , and t are understood for Σ_T , W , and S):

$$\rho(t) \equiv \frac{\int_{\text{reactor}} dV \int_0^\infty dE \int d\Omega W [-v\Omega \bullet \nabla S - v\Sigma_T S + AS + \sum_i X^i(E)GS]}{\int_{\text{reactor}} dV \int_0^\infty dE \int d\Omega W \sum_i X^i(E)GS}, \quad (78)$$

where $X^i(E)$ is the total fission spectrum for isotope i :

$$X^i(E) \equiv X_p^i(E) \{1 - \beta^i\} + \sum_{j=1}^6 X_{dj}(E) \beta_j^i. \quad (79)$$

Now note, if both the numerator and denominator of Eq. 78 are multiplied by the amplitude function $T(t)$, and W is taken as 1, then the numerator is the total neutron production rate minus loss rate for the reactor. (When the first term in the bracket in the numerator is integrated, and Gauss’ theorem is applied, it yields the total neutron leakage rate from the reactor.) Similarly, the denominator is the total neutron production rate. Reactivity is a dimensionless parameter whether or not W is unity. If it is zero, the reactor is critical, if negative, subcritical and if positive, supercritical.

β_j in Eq. 72 is the *effective delayed neutron fraction* for the j th precursor group, and β in Eq. 71 is the sum of the β_j ’s:

$$\beta_j \equiv \frac{\int_{\text{reactor}} dV \int_0^\infty dE \int d\Omega W \sum_i X^i(E) \beta_j^i GS}{\int_{\text{reactor}} dV \int_0^\infty dE \int d\Omega W \sum_i X^i(E) GS}. \quad (80)$$

As with the expression for reactivity, multiplying the numerator and denominator of Eq. 80 by $T(t)$ and letting $W = 1$, one sees that in this case β_j is the fraction of total fission neutrons produced in the reactor by precursor group j .

The parameter Λ is the *prompt neutron lifetime* and is defined as:

$$\Lambda \equiv \frac{\int_{\text{reactor}} dV \int_0^\infty dE \int d\Omega WS}{\int_{\text{reactor}} dV \int_0^\infty dE \int d\Omega W \sum_i X^i(E) GS}. \quad (81)$$

Again multiplying the numerator and denominator by $T(t)$ and taking $W = 1$, one sees that Λ equals the number of neutrons in the reactor divided by the rate of neutron production by fission. If the reactor is critical, Eq. 81 has the same form as the “fundamental equation of radioactive decay,” Eq. 11, and Λ can be thought of as the “mean lifetime” of a neutron born into the reactor. In the point kinetics equations, Λ is the mean *prompt neutron lifetime*, and the timing

of the appearance of delayed neutrons is treated explicitly through the behavior of their precursor, $C_j(t)$ $j = 1, 2, \dots, I$ (usually = 6).

The derivation of the point kinetics equations directly from the time-dependent neutron transport equation has been presented here to provide perspective on approximations that are normally made to make the generation of the point kinetics parameters (ρ , β , and Λ) practical. If Eq. 51 and its *auxiliary* equations could be readily solved for $N(\mathbf{r}, E, \Omega, t)$, there would be no need for the point kinetics equations. As it turns out, however, “practical” approximations follow from the approaches described in the previous section for solving the time-independent transport equation. Henry in *Nuclear-Reactor Analysis* derives the point kinetics equations starting with the diffusion approximation (with energy a continuous variable). One could just as well start with a few group diffusion approximation which would provide a shape function (a vector) and from the adjoint of the few group diffusion equation, a weight function (also a vector). The advantage of using an adjoint weight function, irrespective of the approximation to the transport equation one starts from, is in calculating reactivity, $\rho(t)$. In transients of interest in design, $\rho(t)$ is the driver. It reflects changes in cross sections with time due to variations in temperature (coolant density, Doppler effects) and configuration (control rod motion). In whatever form Eq. 78 takes, given the neutron transport approximation used, if a changing cross section is represented as a starting value plus a time-varying small delta, $\delta\Sigma(\mathbf{r}, E, t)$, and the shape function is represented as a time-independent function (e.g., from the initial neutron density of the reactor, the same problem that generates the adjoint) plus a time-dependent small delta, $\delta S(\mathbf{r}, E, t)$, then the resulting calculation of $\rho(t)$ will to first order in deltas only depend on $\delta\Sigma(t)$ ’s. Higher order terms can be ignored, and one does not need to calculate a time-dependent shape function. This is a classic perturbation problem. Henry (chapter 7) provides a detailed derivation.

The time dependence of $\beta_j(t)$ and $\Lambda(t)$, Eq. 80 and Eq. 81, as used in the point kinetics equations can be ignored in most applications. Measured values of β (and the β_j^i from which it is summed) can be used directly (Table 4). If adjoint weighting is used, the β ’s

will be a bit larger than the physical β ’s in a thermal reactor due to the increased “importance” of delayed neutrons with their lower initial energies (relative to prompt neutrons). Prompt neutron lifetimes primarily depend on the reactor type; for thermal reactors they are on the order of $\sim 10^{-3}$ s, and for fast reactors as short as 10^{-7} s. They can be measured in zero power reactor mock-up experiments as ratios with β and ρ , or calculated directly from equation Eq. 81 with the approximations for W and S used to obtain ρ . A highly accurate calculation of Λ can be made with a Monte Carlo simulation where neutrons are introduced from the prompt neutron energy spectrum with an $S(t = 0)$ spatial distribution. Each history would be timed and terminated with neutron absorption (capture plus fission) or leakage. The “times” will yield the mean and standard deviation of the prompt neutron lifetime.

One further note on reactivity, if a perturbation of a critical reactor configuration can be viewed as nearly instantaneous, that is, a step change, then a good estimate of reactivity addition or subtraction can be found by solving the eigenvalue (time independent) problem for the perturbed reactor:

$$\rho = 1 - 1/k, \quad (82)$$

and if the initial reactor configuration is subcritical then the reactivity addition from a “step” perturbation can be found by performing two eigenvalue problems, the perturbed case as before, and one for the initial subcritical configuration (ignoring any nonfission source);

$$\rho = 1/k_0 - 1/k, \quad (83)$$

where k_0 (< 1) is the initial subcritical eigenvalue. As reactivity is a dimensionless ratio, it is often given as a percentage or in units of β (as defined by Eq. 80). In the latter case, the “units” are traditionally dollars and cents. If ρ equals β , the reactivity addition is 1 dollar; if ρ equals 0.5β the reactivity addition is 50 cents. If a dollar of reactivity is added to a critical reactor, it is said to be prompt critical (critical on prompt neutrons alone) and delayed neutrons will not mitigate the resulting transient, a condition obviously to be avoided.

The motivation for the description of point kinetics provided here is best provided by Henry:

- “Since mean free paths are fairly long and since the lifetimes of neutrons in a reactor are quite short, the effects of local perturbations on” $N(\mathbf{r}, E, \Omega, t)$ “will quickly spread throughout a reactor. The immediate consequences of perturbing a reactor locally (for example by changing a control rod slightly) is thus a readjustment of the shape of the” neutron density. In many cases this readjustment is slight and is completed in a few milliseconds; after that the readjusted shape rises or falls as a whole depending on whether the initial perturbation increases or decreases k_{eff} . For reactors in which transients proceed in this manner, merely being able to predict the change in *level* of the neutron density “is sufficient to permit a very accurate prediction of the consequences of perturbation.”

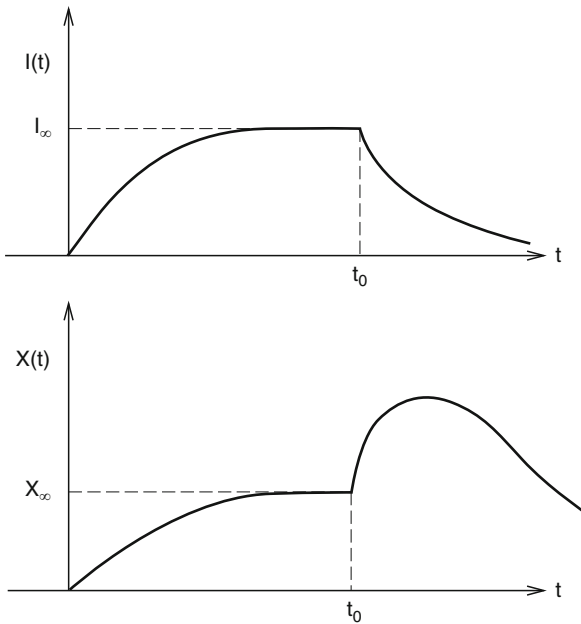
With today’s computer capabilities, solving point kinetics problems is not a great challenge, even with time-dependent reactivity reflecting feedback from power changes in the reactor. Henry and Duderstadt provide descriptions of applicable calculational methods (development of which inspired great ingenuity in the past). In any case, to quote Henry again, point kinetics solutions provide “very accurate predictions of the consequences of (reactor) perturbations.” Thus, their utility in assuring that a reactor design satisfies fundamental transient requirements. Under normal operating conditions, a reactor must be inherently stable, that is, self-limiting. Reactivity must be reduced with increased temperature; that is, with reduced coolant density, increased mean thermal neutron energy, and Doppler broadening of resonance cross sections. These phenomena are dependent on reactor type. Clearly change in thermal neutron spectra is unimportant in a fast reactor. However, if coolant density decrease results in voiding, reactivity will dramatically decrease in a thermal reactor, but in a liquid metal cooled fast reactor increased leakage must outweigh a higher energy neutron spectrum (and an increased fission to capture ratio in fuel) to assure a negative reactivity effect. Also movement of control rods must reduce reactivity when reactor shutdown is desired. Some movable poisons, power shimming control rods, could be included in a design to flatten (make more uniform) the power distribution throughout life.

(Power flattening can reduce coolant pumping power requirements. As coolant flow must meet the heat removal needs of the hottest region of the reactor, minimizing excess flow to cooler regions increases overall power plant efficiency.) But, one must assure that the operating strategy for using such rods does not compromise the speed of reactor shutdown when it is required to deal with an accident condition.

Point kinetics models can also aid in assessing *Xenon override* requirements for thermal reactors. Xe^{135} with its extremely large thermal neutron capture cross section ($\sigma_c = 2.7 \times 10^6$ b at $E = 0.023$ eV, for comparison $\sigma_f^{U235} = 577$ b at 0.023 eV) is the most important fission product poison. It is produced directly from fission and by decay of its precursor I^{135} (which is a direct fission product and has short-lived precursors, $\text{Sb}^{135} \rightarrow \text{Te}^{135} \rightarrow \text{I}^{135}$). Both Xe^{135} and I^{135} have half-lives measured in hours ($T_{1/2}^{\text{Xe}} = 9.14$ h, $T_{1/2}^{\text{I}} = 6.58$ h). So, a point kinetics reactor model will show that when a reactor is started up, Xe^{135} and I^{135} build up to equilibrium levels in about 30 h. Their levels, inventories, will depend on power level, that is, neutron density, and the reactor design must have enough excess reactivity (e.g., control rods that can be withdrawn, or for a PWR, a soluble poison in the coolant whose concentration can be reduced) to “override” the negative reactivity perturbation of neutron capture by Xe^{135} (iodine neutron capture can be ignored). In addition, when a thermal reactor is shut down, Xe^{135} builds up as loss by neutron absorption stops and decay of I^{135} continues. The Xe^{135} inventory peaks in about 10 h to approximately three times its equilibrium level and subsequently decays. It is back to its equilibrium level in ~ 40 h (see Fig. 8).

Clearly additional “excess reactivity” is required to deal with a post-shutdown Xenon transient. Eventually, not being able to provide (and control) this excess reactivity can limit the useful lifetime of a thermal reactor design. As a historical aside, it was John Wheeler [3] who recognized the role that Xe^{135} and I^{135} could play in the operation of a thermal reactor. He explained the initial difficulties in operating the first plutonium production reactor at the Hanford Washington site.

While point kinetics can deal with most reactor design transient requirements, there are instances where significant spatial effects must be accounted



Fission Reactor Physics. Figure 8

Schematic of I^{135} and Xe^{135} inventories following an initial reactor start-up and subsequent shutdown after equilibrium levels have been reached (i.e., at t_0). (Duderstadt)

for. In the normal operation of a large thermal reactor, there is the potential for spatial oscillations of Xenon concentrations, and therefore neutron density and power. These oscillations may not affect criticality and might only be observed by the reactor instrumentation system's ability to measure power distributions in core. The period of these oscillations would be many hours and thus limits on coolant channel performance and fuel temperature could be compromised for extended time intervals. Space-time calculations of neutron density have to be carried out to determine if a particular design has a propensity for these oscillations. If the design is not inherently stable, its instrumentation must provide detection and an operating strategy must be devised to suppress any oscillation initiation. With modern computing capability, few group diffusion approximations to the time-dependent transport equation, Eq. 51 (with explicit spatial mesh or nodal methods), can deal with Xenon oscillation evaluations [22].

There are certain reactor plant accident scenarios, which also require space-time calculations. To deal

with these, nodal methods, as described in section “Neutron Distributions”, have been incorporated in safety analysis programs (see for example [23]), which model a full power plant; the reactor, its neutronics, fluid mechanics and structural integrity; and the balance of the plant, instrumentation, coolant/working fluid to power conversion, and safety systems (containment, emergency coolant injection and power supply). There are two classes of accidents where space-time effects in the reactor are important. First is “rod ejection,” where a single control rod or group of control rods is rapidly withdrawn with a large reactivity addition and neutron distribution change, in no way a small perturbation in the point kinetics sense. Second is a “cold water accident” applicable to a PWR. In this case, in a plant with multiple piping “loops” carrying coolant to the reactor, if one of the loops is not functioning so that water in the loop has cooled below the normal inlet temperature of the reactor, and the loop is reactivated (its pump turned on and isolation valves (if any) opened), while the reactor is critical, there can be large asymmetric reactivity insertion, again, this is not a small perturbation problem.

Finally, there is an additional aspect of reactor neutron time variation, which has interesting *Physics*. The transport equation, derived in section “Neutron Distributions”, is for the *mean* neutron density, $N(\mathbf{r}, E, \boldsymbol{\Omega}, t)$, but clearly as neutron interactions and production, and fission product decay are inherently probabilistic, there are fluctuations in the neutron population in a reactor (and as there are in delayed neutron precursor populations). These fluctuations were recognized early on in the Manhattan Project and are also referred to as neutron noise [24]. There are several approaches to modeling of the phenomena. The most fundamental is based on the derivation of the neutron transport equation from the quantum Liouville equation (Osborn and Yip [25]). This derivation is extended to produce an equation for the neutron doublet density, $N^{NN}(\mathbf{r}, E, \boldsymbol{\Omega}, \mathbf{r}', E', \boldsymbol{\Omega}', t)$, the expected number of neutrons in d^3r about $\mathbf{r}$, with energies in dE about E , going in the solid angle $d\boldsymbol{\Omega}$ about $\boldsymbol{\Omega}$ times the expected number in d^3r about $\mathbf{r}'$, and so on for E' and $\boldsymbol{\Omega}'$, where as in the transport equation neutron–neutron collisions are ignored. Additional equations for neutron–precursor and precursor–precursor doublet densities are produced to complete the set of equations needed to

solve for N^{NN} . One also needs equations for the “mean” (in Osborn’s nomenclature, singlet) neutron and precursor densities, which were derived in section “[Neutron Distributions](#)” (Eq. 51 and Eq. 52). Being able to solve for N^{NN} is, however, not sufficient to predict the results of neutron “noise” experiments in a reactor. Equations for doublet and singlet densities for events (D) in detectors (e.g., pulse or continuous currents in an ionization chamber) are also needed. With N^{DD} and N^D one can predict variance to mean ratios of counts or the power spectral density of the current in a single detector, and the cross power spectral density of currents in two detectors [26]. These experiments are usually performed in a reactor in a steady-state condition (in the mean of course) at zero power, critical or slightly subcritical, to obtain estimates of point kinetics parameters. These experiments have the advantage of verifying expected kinetic performance without putting the reactor into a transient. They are performed at low neutron levels because as power is increased fluctuations become negligible relative to the mean. Detector noise measurements (psd and cpsd) are sometimes made at power in operating reactors to monitor for unplanned mechanical motion, or loose parts. Modeling for these measurements is deterministic.

Fluctuations in neutron populations must be considered in developing initial start-up procedures for a newly constructed reactor (In a reactor design which incorporates fuel (including fission products and transuranics) from a previously operated reactor, natural source levels will most likely be high enough to allow “fluctuations” to be ignored.). The mean neutron density before start-up depends on an external source of neutrons, $S(\mathbf{r}, E)$ (not from neutron reactions, Eq. 50) which, as part of the design, could be adjacent or internal to the reactor. At start-up, the reactor is subcritical ($k < 1$) and the mean neutron population, in a point reactor sense, is $N = \Lambda S(k/(1-k))$. In outline, the steps to bring the reactor critical are to pull control rods up (down in most BWR designs) from their fully inserted position so as to insert some precalculated amount of reactivity and then wait for a new steady neutron level to be achieved. The subcritical neutron level is monitored by the reactor’s source range detectors. A power reactor is instrumented with detectors for the full range of expected neutron levels. These pull-and-wait steps are repeated until the reactor is slightly

super critical, and thus the neutron level is observed to be on a continuously increasing, but easily controlled, trajectory. A pull-and-wait procedure needs to account for neutron level fluctuations because if the observed level, due to a minimizing fluctuation, is below the expected (mean) level when the reactor is actually close to its critical configuration, then the next “pull” might produce an unacceptable rapidly increasing trajectory [27]. The simplest way to avert a problem with a pull-and-wait procedure is to assure that the sources provided in the design (In a reactor design which incorporates fuel (including fission products and transuranics) from a previously operated reactor, natural source levels will most likely be high enough to allow “fluctuations” to be ignored.) are strong enough to render subcritical neutron fluctuations negligible. Of course, one has to understand neutron level fluctuations to make this assessment [28].

Future Directions

Fission reactor development, since its inception, has progressed with advances in computing capability. Early on, analog computers were used for transient analyses, but digital computers have been the primary tool. Design codes have been adapted to advancing digital technology; scalar processing, then vector processing and now massively parallel processing. This is an active field today and should continue to be so, particularly as the drivers for improve computer technology are universal. In section “[Neutron Distributions](#)” the two approaches to solving the neutron transport equation, Monte Carlo simulation and analytic methods (differencing variables and solving the resulting matrix equations) were described. Parallel computing would appear to be well suited to Monte Carlo as independent histories can be run on the various (1,000’s of) processors simultaneously. There is, of course, the need to provide the reactor configuration (geometry, nuclide inventories, and cross sections) and the Monte Carlo code itself to each processor that runs a history. This challenge is being accepted with considerable success as exemplified by the accomplishments of the Los Alamos National Laboratory group working on the MCNP code [15] and the joint effort at the Knolls and Bettis Atomic Power Laboratories on the MC21 code [29]. Advocates of the analytic approach

have, however, not accepted the ultimate triumph of Monte Carlo. This is clear in the work of a group at the Argonne National Laboratory, which has modestly named their multigroup, discrete ordinate code UNIC, for Ultimate Neutronic Investigation Code [30]. They are demonstrating impressive results for fast reactor designs. Competition in supercomputer development and in attendant codes for nuclear reactor design bodes well for better products in the future.

Physicists in their efforts to understand the atomic nucleus have made myriad measurements and only partially by design these have included the neutron and gamma cross sections, and fission product yields (and their decay mechanisms) needed for the development of fission reactors. Today, work on this reactor-related data is focused on establishing well-founded uncertainty measures. The Cross Section Evaluation Working Group of the National Nuclear Data Center refers to this effort as covariance evaluation [31]. This is particularly appropriate, for as discussed above, one expects calculational methods to improve with computer power and code development. Thus, in assigning error bounds in design, the uncertainty in basic data will become more important relative to the contribution of calculational error (e.g., Monte Carlo statistics or differencing and convergence error in analytic methods). More well-founded and hopefully smaller design error bounds can obviously be taken advantage of in future reactor development. Improving error bounds is also consistent with the approach to overall power plant safety analysis being fostered by many of the world's nuclear regulatory agencies. They favor best estimate analyses plus the assignment of rigorously defined uncertainty factors for various classes of accident conditions. Reactor design error is only a contributor to a safety analysis "uncertainty factor," but for power plant technology to advance, "reactor design" must do its part.

Much of the research for a next generation of fission reactor power plants is focused on higher operating temperatures. Today's thermal reactor plants, PWRs, BWRs, and CANDU (heavy water moderated) have outlet reactor coolant temperatures, T_H , of $\sim 600^\circ\text{F}$, and thus thermal (Rankine cycle) efficiencies in the low 30%. Raising T_H would increase cycle efficiency and lower fuel cost, and provide high-temperature process heat (possibly for a catalytic hydrogen production process). Higher operating temperatures in a water

(or heavy water) environment present core and structural materials challenges. There likely will be a need for additional resonance cross-section data for some additives (e.g., Manganese and Chrome) to new high-temperature materials. Fast reactors (liquid metal or gas cooled) operate at much higher temperatures than thermal reactors but also require much higher fissile inventories to attain criticality. Experience with design and operation of these reactors is limited (especially for gas cooled reactors) compared to thermal reactors. Their future development with emphasis on the burning of unwanted transuranics as well their traditional mission of efficient conversion of fertile isotopes (U^{238} and Th^{232}) to fissile isotopes (Pu^{239} , Pu^{241} and U^{233}) will stimulate some cross-section work. But, both thermal and fast reactor development will most likely benefit more from advances in branches of physics other than Nuclear, particularly Condensed Matter and Fluid Mechanics. The need for nuclear power, both economic and environmental (if they can be separated?), will drive fission reactor development. While Uranium and Thorium are abundant in the earth's crust, power demand will push reactor development to most efficiently exploit the highest grade ores, which will likely lead to a mix of fast and thermal reactors. In any case, physics as well as engineering will play key roles in this development.

Bibliography

Primary Literature

1. Wheeler JA (1967) Mechanism of fission. *Phys Today* 20(11): 49–52
2. Bohr N, Wheeler JA (1939) The mechanism of nuclear fission. *Phys Rev* 56:426
3. Ford KW (2009) Wheeler's work on particles, nuclei, and weapons. *Phys Today* 62(4):29
4. Fermi E, Szilard L (1955) Neutronic reactor. US Patent 2,708,656, 17 May 1955
5. Einstein A (1905) On the electrodynamics of moving bodies. *Ann Phys* 17:891–921
6. Frankel S, Metropolis N (1947) Calculations in the liquid-drop model of fission. *Phys Rev* 72:914
7. Kinsey R (1979) Compiler "ENDF/B Summary Doc." BNL-NCS-17541(ENDF-201), 3rd edn. ENDF/B-V, Brookhaven National Laboratory, Upton
8. England TR, Wilson WB, Stamatelatos MG (1976) Fission product data for thermal reactors. EPRI-NP-356, Los Alamos Scientific Lab., New Mexico

9. Parks DE, Nelkin MS, Wikner NF, Beyster JR (1970) Slow neutron scattering and thermalization with reactor applications. Benjamin, New York
10. MacFarlane RE (1994) The NJOY nuclear data processing system, Version 91, report #LA-12740-M, Los Alamos National Laboratory, New Mexico
11. Breit G, Wigner E (1936) Capture of slow neutrons. *Phys Rev* 49:519
12. Beckurts KH, Wirtz K (1964) Neutron physics. Springer, Berlin
13. Larson NM (2006) Updated user's guide for SAMMY: multilevel R-matrix fits to neutron data using Bayes' equations, ORNL/TM-9179/R7. Oak Ridge National Lab, OakRidge
14. Spanier J, Gelbard EM (1969, 2008) Monte carlo principles and neutron transport problems, Dover, New York, Addison-Wesley, Reading
15. X-5 Monte Carlo Team (2003) MCNP – A general Monte Carlo N-Particle transport code, Version 5, LA-UR-03-1987, Los Alamos National Laboratory, New Mexico
16. Ondis LA II, Tyburski LJ, Moskowitz BS (2000) RCP01 – a Monte Carlo program for solving neutron and photon transport problems in three dimensional geometry with detailed energy description and depletion capability, B-TM-1638. Bettis Atomic Power Laboratory, West Mifflin
17. TOP 500 Super Computing Sites, www.Top500.org
18. Lathrop KA (1972) Discrete-ordinates methods for the numerical solution of the transport equation. *Reactor Technol* 15:107
19. Lathrop KA (1971) Remedies for ray effects. *Nucl Sci Eng* 45(3):355–68
20. Varga RS (2000) Matrix iterative analysis, 2 Revised and Expandedth edn. Springer, Heidelberg
21. Henry A, Dias A, Frances W, Parlos A, Tanker E, Tanker Z (1986) Continued development of nodal methods for nuclear reactor analysis, MIT EL 86-002. Massachusetts Institute of Technology, Cambridge
22. Buslik AJ, Weinreich WA (1967) Variational calculation of complex natural modes of xenon oscillation, WAPD –TM-673 (LWB-LSBR Development Program). Bettis Atomic Power Laboratory, West Mifflin
23. Idaho National Laboratory (2009) RELAP5-D, www.INL.gov
24. Thie JA (1963) Reactor noise (an AEC monograph). Rowman and Littlefield, New York
25. Osborn RK, Yip S (1966) Foundations of neutron transport theory, Monograph series on nuclear technology. Gordon and Breach, New York
26. Natelson M, Osborn RK, Shure S (1966) Space and energy effects in reactor fluctuation experiments. *J Nucl Energy Parts A/B* 20(7):557–585
27. Hurwitz H Jr, MacMillan DB, Smith JH, Storm ML (1963) Kinetics of low source reactor startups, part I and II. *Nucl Sci Eng* 15:166
28. Clark WG, Harris DR, Natelson M, Walter JF (1968) Variances and covariances of neutron and precursor populations in time-varying reactors. *Nucl Sci Eng* 31:440–457
29. Sutton TM et al, Knolls Atomic Power Lab., Griesheimer DP et al, Bettis Atomic Power Lab. (2007) The MC21 Monte Carlo Code, LM-06K144, Proceedings (on CD-ROM) of the joint international topical meeting on M&C and supercomputing applications, Monterey
30. Palmiotti G et al, Argonne National Lab. (April 2007) UNIC Ultimate Neutronic Investigation Code, Proceeding (on CD-ROM) of the joint international topical meeting on M&C and supercomputing applications, Monterey
31. Herman M et al (2008) Covariance evaluation methodology for neutron cross sections, BNL-81525-2008. Brookhaven National Lab, Upton
32. Belle J, Berman RM (eds) (1984) Chapter 2, Natelson M, Nuclear properties of the thorium fuel cycle. In: Thorium dioxide: properties and nuclear applications, Naval Reactors Office, U. S. Department of Energy, Govt. Printing Office, Washington, DC
33. Katcoff S (1958) Fission–Product yields from U, Th and Pu. *Nucleonics* 16(4):78

Books and Reviews

- Duderstadt JJ, Hamilton LJ (1976) Nuclear reactor analysis. Wiley, New York
- Evans RD (1955) The atomic nucleus. McGraw-Hill, New York
- Henry AF (1975) Nuclear-reactor analysis. MIT Press, Cambridge
- Kaplan I (1962) Nuclear physics, 2nd edn. Addison-Wesley, Reading
- Krane KS (1988) Introductory nuclear physics. Wiley, New York
- Mermin ND (2005) It's about time, understanding Einstein's relativity. Princeton University Press, Princeton

Fly Ash

MEI-IN MELISSA CHOU

Illinois State Geological Survey-Prairie Research Institute and Department of Natural Resources and Environmental Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Fly Ash Production and Rate of Utilization
 Fly Ash Management
 Fly Ash Properties and Classification
 Applications
 Environmental Impacts and Considerations
 Future Directions
 Bibliography

Glossary

Cementitious Said of material that has self-cementing properties in the presence of water alone.

Cenospheres Known as “microspheres” that are small, lightweight, inert, hollow spheres filled with air and/or gases, comprised mainly of silica and alumina, produced from coal combustion process like fly ash, and used widely mainly as high-quality, lightweight fillers.

Coal A sedimentary rock of organic origin, consisting mainly of carbonized plant remains and containing some other elements, such as sulfur, hydrogen, and nitrogen, and mineral impurities.

Coal combustion by-products (CCPs) Residues generated from the combustion of pulverized coal, which include fly ash, bottom ash, boiler slag, and flue gas desulfurization (FGD) materials.

Coal mineral matter Material in coal derived from minerals present in the original plant materials that formed the coal, or from external sources such as sediments deposited with the plant remains or precipitates from interstitial water. This mineral matter converts to ash during coal combustion.

Concrete A construction material produced from a mixture consisting primarily of aggregates (sand and crushed stone or gravel), Portland cement, and water. Certain coal ashes can be used as a replacement for a portion of the Portland cement.

Pozzolan Siliceous or aluminosilicate materials in fine powdery form that, in the presence of water, will chemically react with the calcium hydroxide released by hydration of Portland cement to form compounds possessing cementitious properties.

Supplementary cementing material A material that, when used in conjunction with Portland cement, contributes to the properties of hardened cement through its hydraulic or pozzolanic activity, or both.

Definition of the Subject

Fly ash consists of the inorganic matter within coal that has been melted at high temperature during coal combustion, solidified while suspended in the flue gases, and collected by electrostatic precipitators or other devices before the flue gases are released into the atmosphere. Fly ash is the lighter, less dense part of coal ash, consisting of small spherical particles having

pozzolanic or some cementitious properties. It is distinguished from the heavier, denser part of coal ash, the bottom ash, which is collected at the bottom of the boilers and consists of coarse nonspherical particles that lack pozzolanic/cementitious properties. Fly ash, also known as pulverized fuel ash (PFA) because of its fine powdery form and pozzolanic and cementitious nature, has many established applications in the production of concrete-related construction and pavement materials.

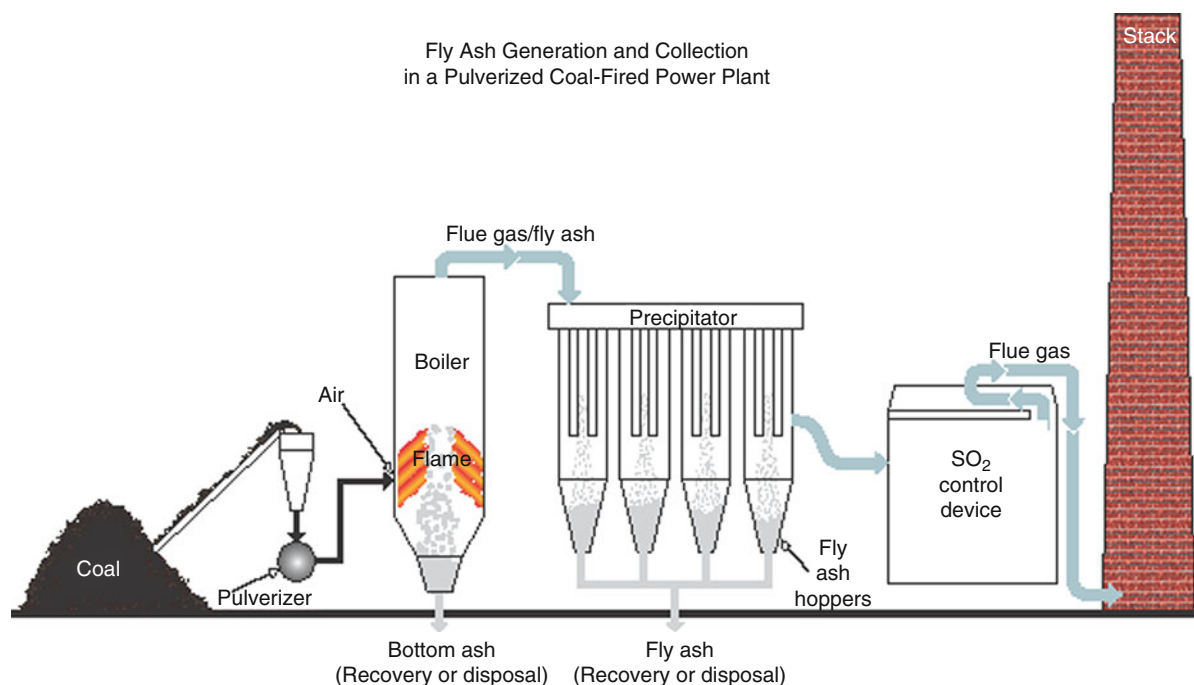
Introduction

The pulverized coal firing system for electricity generation was developed in early twentieth century in the USA, and since then, the system has been adopted in more than 50 countries. As pulverized-coal-fired generating stations were constructed worldwide, increasing quantities of coal combustion by-products became available. Fly ash is the most abundant coal combustion by-product and the amount generated has increased dramatically during the past few decades. It has become apparent that simply discarding such large quantities of this material would be environmentally unacceptable and methods for its reuse have been developed for years, especially in Europe, the USA, India, and China. Nevertheless, there is a great variation in usage among countries, from total reuse to almost complete disposal. Efficient, environmentally sound, and cost-effective handling of fly ash is a continuing challenge for major fly ash producers, marketers, and regulators worldwide. This entry provides useful resources and a description of the generation, specifications, management, and utilization of fly ash produced by pulverized-coal-fired power plants.

Fly Ash Production and Rate of Utilization

A schematic flow diagram of fly ash production in a typical pulverized-coal-fired power plant equipped with an electrostatic precipitator for fly ash collection is shown in [Fig. 1](#).

In a coal-fired power plant, pulverized coal is burned in a boiler under high temperature and the heat generated converts water to steam for electricity generation. The noncombustible mineral matter in coal (typically 10–25 wt% in US coals and 30–50 wt% in the coals of India) is converted to coal ash. Fly ash,



Fly Ash. Figure 1

Fly ash generation in a pulverized-coal-fired power plant equipped with a dry-bottom boiler, an electrostatic precipitator, and a flue gas desulfurization (FGD) device

the lighter part of coal ash, is suspended in the flue gas and is collected before release of flue gas into atmosphere. Fly ash collection systems include electrostatic precipitators, baghouses, fabric filters, or mechanical devices such as cyclones. The electrostatic precipitator is the most common and efficient of these; it can remove about 99.9% of the fly ash in the flue gas. Using a dry-bottom pulverized-coal-fired boiler, fly ash accounts for 75–85% of the total coal ash produced [1] and the remainder is the heavier, denser part, the bottom ash, which is collected at the bottom of the boilers (Fig. 1). The solid waste materials derived from coal combustion are known as “combustion residues” or “coal combustion by-products (CCPs).” If a pulverized-coal-fired power plant is equipped with a flue gas desulfurization (FGD) system for burning high-sulfur coal, another type of CCP, known as FGD residues, is also produced. The amount of all CCPs being generated has increased dramatically around the world during the past few decades as the number of coal-fired power plants increased and as pollution regulations forced plants to reduce

their emissions. However, fly ash is the most abundant among these CCPs.

Organizations have been created in many countries, usually with the support, or collaboration of their governments, for the purpose of minimizing landfill and maximizing the utilization of CCPs. They provide information and assistance to those who are interested in developing or using coal combustion by-products in value-added applications. These organizations also collect statistical data, and analyze trends in the production and utilization of CCPs and their disposal practices. These organizations include, but are not limited to, the American Coal Ash Association (ACAA), the European Coal Combustion Products Association (ECOBA), the Ash Development Association of Australia (ADAA), the United Kingdom Quality Ash Association (UKQAA), and the Fly Ash Utilization Program (FAUP) in India. For a more comprehensive list see “Useful Resources” at the end of this entry.

ACAA compiled data on production and usage of CCPs in the USA as early as 1966. They indicate that the fly ash production rate in the USA increased from

about 59.4 million tons in 1996 to about 68.1 million tons in 2001 and to 72.4 million tons in 2008. ACAA's figures use short ton (1 short ton=2,000 lb=0.9072 metric tons). The amount of fly ash recycled into commercial products also increased from about 22% in 1996, to about 32% in 2001, to about 42% in 2008. More than half of the remaining fly ash produced in 2008 was still deposited in landfills or storage ponds [2] at a cost to the utility companies and thus to the consumers.

Typical coals in India have ash contents ranging from 30 to 50 wt% [3]. During 1999–2000, the utility industry in India produced about 70 million tons of coal ash per year, but only about 10% of the ash was used in ash dike construction and land filling, about 3% was used in other construction products, and the remainder was disposed of in landfill [4]. In other more developed countries, 80% or more of the fly ash generated was used in manufacturing concrete, bricks, and cellular concrete blocks, in soil-stabilization treatment and land fill applications, in agriculture, in recovery of metals and cenospheres, and in dam construction.

Realizing coal ash posed a significant environmental problem, in 1999, the government of India set up the “Flyash Mission” in New Delhi under its Department of Science & Technology with a goal of coordinating ash utilization efforts and aiming at 100% fly ash utilization [3]. A law was approved by Indian government in 1999 [5] requiring every construction agency engaged in the construction of buildings within a radius of 50 km (amended to 100 km in 2003 [6]) from coal-fired power plants to use only fly ash-containing construction products, and those manufacturing fly ash construction products should contain at least 25 wt% fly ash. During 2004–2005, when the utility industry in India generated about 112 million tons of fly ash per year, the FAUP (Fly Ash Utilization Program) in India has succeeded in increasing fly ash utilization from 1 million tons per year in 1994 to about 42 million tons per year in 2005 [7]. However, the remaining 70 million tons of fly ash still ended up being deposited in ash ponds or lagoons that covered more than 26,000 ac of arable land [8].

In China, there are incentive measures to encourage utilization of fly ash, especially in the concrete and cement industry. For example, the company that produces cement products with more than 30% of fly ash will

receive tax deduction. Some areas of China such as Beijing, Shanghai, and Hubei, recycle 100% of their fly ash. The national figures from China Energy Development report indicated that total fly ash production in China increased from 148 million tons in 2000, to 229 million tons in 2004, and is predicted to reach 378 million tons in 2010. The fly ash usage rate also increased from 58% in 2000 to 68% in 2004. Similar to other major fly ash-producing countries, the remaining huge amount ended up in ash ponds [9].

Many individual scientific papers have described the production and usage rates of fly ash from around the world. Dhadse and coworkers published a review paper on fly ash characterization, utilization, and government initiatives in India [10]. In their paper, the fly ash production and usage rates around 2005 in various countries were compared (Table 1), with data from TIFAC (Technology Information, Forecasting and Assessment Council) in India <http://www.tifac.org.in>. As indicated in Table 1, the US figures differ from those reported by the ACAA, and the figures for China differ widely from those in the China Energy Development report described earlier in this section [9]. The ACAA report indicates that in 2005, the USA produced about 71.1 million (short) tons of fly ash and only about 41% of it was used. Nevertheless, Table 1 provides

Fly Ash. Table 1 Fly ash generation and utilization in different countries (printed with permission of the Journal of Scientific & Industrial Research [10])

Country	Annual ash production, MT	Ash utilization, %
India	112	38
China	100	45
USA	75	65
Germany	40	85
UK	15	50
Australia	10	85
Canada	6	75
France	3	85
Denmark	2	100
Italy	2	100
Netherlands	2	100

a reasonably accurate worldwide comparison in fly ash production and usage. These data indicate that fly ash production in different countries varied from 2 to 112 million tons with the major producers being India, China, and the USA. The fly ash utilization rate varied widely from 38% to 100%, with the total fly ash being recycled in the countries of Denmark, Italy, and the Netherlands.

Worldwide production and consumption of coal for energy has increased dramatically, especially in recent years, in countries such as India and China. This is because the total world reserves for coal (250–1,500 years at current consumption rates) are the most abundant, cheap, and easily produced energy resource, compared to the reserves for natural gas (70–120 years) and crude oil (40–60 years) [11]. The cost of using coal for electricity generation may become more expensive in the future when regulations for green house gas emissions are implemented. Currently, coal provides more than half of the US electricity supply and will continue to be the fuel of choice for worldwide power generation for a couple decades to come [11]. Therefore, managing coal fly ash in an efficient, environmentally sound, and cost-effective manner will remain a continuing worldwide challenge.

Fly Ash Management

Figure 1 shows fly ash produced in a typical pulverized-coal-fired utility boiler and collected by an electrostatic precipitator. There are two commonly used methods for removing the fine powdery fly ash from the fly ash hopper of the precipitator – a wet method or a dry method [12–15].

In the wet method, water is used to flush the fly ash out of the hoppers and the ash slurry is pumped through pipes to large, man-made settling ponds in the vicinity of the power plant. This method was widely used because it was the cheapest method for removing and storing fly ash. Utility companies that use an ash pond for fly ash disposal have three basic options available when the ponds are full: (1) constructing another ash pond nearby; (2) excavating the fly ash from the pond and transporting it to a land fill; and (3) raising the edge of the existing pond by constructing a dike to allow disposal of more fly ash in the pond.

In the dry method, vacuum suction or high-pressure compressed air is used to fluidize and transport the dry fly ash to storage silos, from which it can be either reused or disposed of in a dry or wet form. Most silos are equipped with fluidizing capability for easy discharge of the fly ash through their bottom outlets. Fly ash in a dry form can be directly discharged into enclosed containers, tanks, or wagons [13]. Typically, a little water is added to the dry fly ash to allow stockpiling or transporting in a conditioned form (about 15 to 30% moisture) in trucks or rail cars. The main advantage to conditioning the fly ash is to control the dust that may be released during transporting of the material.

In areas where there is an established fly ash market, coal-fired power plants typically would install a dry fly ash handling system to allow beneficial reuse of the fly ash and avoid disposal in impoundments. In areas, with limited or no markets for fly ash, the expense of installing and operating dry collection systems has been difficult to justify if the space needed for ponding or wet stacking is available near the power plant.

Finding a beneficial application is the preferred method for managing fly ash because it provides a way to convert fly ash from a liability to a value-added raw material. Also, the final products containing fly ash, in general, are more durable and of better quality than their conventional counterparts. Environmentally, use of the fly ash can reduce land disposal requirements and reduce the use of other natural resources, such as the lime or clay that the ash replaces. Economically, recycling can reduce the utility's costs associated with fly ash disposal and may increase revenues for both fly ash producers and users by using fly ash to replace higher cost conventional raw materials. The performance benefits may vary depending on the properties of the fly ash and the type of application. For example, the largest single use of fly ash worldwide is in concrete/cement products (see discussion in section "Applications"), and fly ash-based concrete products or pavements have been found to have better engineering properties and longer service life than conventional concrete products.

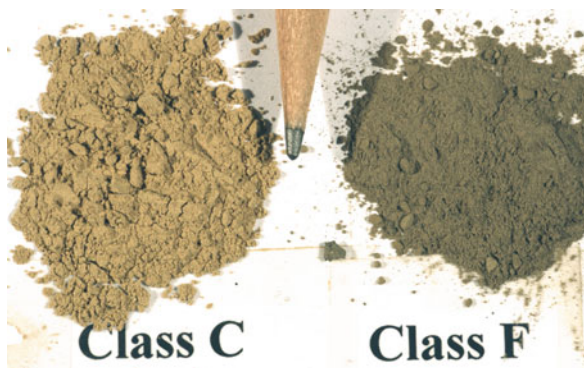
Although power companies produce ash at their coal-fired power plants, most power companies do not handle the disposing or marketing of the ash they produce [16]. Commercial ash marketing companies

operating in many countries have been playing increasingly important roles in fly ash management. These companies are able to maintain a product with consistent chemical and physical characteristics, a factor that is extremely important, especially for concrete/cement product applications. Some companies have facilities for research and development, and facilities for fly ash characterization and processing. They can blend different fly ashes, or remove unburned carbon from a high-carbon fly ash to meet concrete specifications. Some companies provide coal-fired power plants with complete on-site ash handling and management, including the installation of dry fly ash collection systems and the environmental and engineering services. Power plants that use ash marketers to sell their fly ash may employ an ash marketing specialist whose responsibility is to monitor ash generation and quality and interface with ash marketing companies or brokers who are under contract to the utility company.

There are about 40–50 commercial ash marketing companies operating throughout the USA, in all states with an exception of Hawaii [16, 17]. Some commercial ash marketing companies are listed in the “Useful Resources” at the end of this entry.

Fly Ash Properties and Classification

Figure 2 shows photos of small quantities of a bituminous coal fly ash (Class F) and a subbituminous coal fly ash (Class C) highlighting their powdery form.



Fly Ash. Figure 2

Photos of small piles of bituminous coal fly ash (Class F) and subbituminous coal fly ash (Class C). Photos courtesy of the Illinois State Geological Survey

Fly ash consists of fine, powdery particles with diameters ranging from <1 to $150\ \mu\text{m}$. The particles are predominantly spherical in shape and mostly glassy (amorphous, noncrystalline) in nature, because the heat generated during coal combustion is high enough to cause partial melting of many of the minerals present in the coal, and the materials quickly solidify from the molten state while suspended in the flue gas. The color of fly ash can vary from tan to gray to black, depending on the amount of unburned carbon in the fly ash. The darker colors generally indicate greater carbon content. Lignite or subbituminous coal fly ash is typically low in unburned carbon (LOI) content and high in lime (CaO) content (Table 2) and is a light tan to buff color. Most bituminous coal fly ash is typically gray in color, with lighter shades of gray generally indicating a higher quality and lower unburned carbon content fly ash.

Fly ash contains a mixture of crystalline and amorphous materials. Figure 3 shows a typical X-ray spectrogram of a dry and a ponded fly ash sample. The crystalline components include α -quartz (SiO_2) that escaped melting and minerals such as mullite ($3\text{Al}_2\text{O}_3 \bullet 2\text{SiO}_2$), hematite (Fe_2O_3), and magnetite (Fe_3O_4) that formed at high temperature during coal combustion.

Figure 4 shows SEM images, indicating spherical particles, of a bituminous coal fly ash sample at $3,000\times$ magnification (A) and at $6,000\times$ magnification (B). Some fly ashes contain a small proportion of particles that are hollow spheres filled with air/gas that will float on the surface of fly ash ponds. These particles, known as “cenospheres,” are collected and used in some specialized applications (see section “Applications”).

ASTM (American Society for Testing and Materials) C 618 provides specifications for fly ash and raw or calcined natural pozzolan for use as mineral admixtures in Portland cement concrete [18]. It also specifies two classes for fly ash, F and C, on the basis of their chemical composition, which mostly results from burning different types of coal. As shown in Table 2, the chemical composition of fly ash is generally reported in terms of oxides that include silica (SiO_2), alumina (Al_2O_3), and the oxides of calcium (CaO), iron (Fe_2O_3), magnesium (MgO), sulfur (SO_3), sodium (NaO), and potassium (K_2O). Class F fly ash consists of at least 70% of combined SiO_2 , Al_2O_3 , and Fe_2O_3 , whereas, Class C fly ash must have at least 50% of combined SiO_2 , Al_2O_3 , and Fe_2O_3 .

Fly Ash. Table 2 Chemical compositions (wt%) of typical fly ash from burning coal of different ranks and the requirements of the ASTM C 618 classification

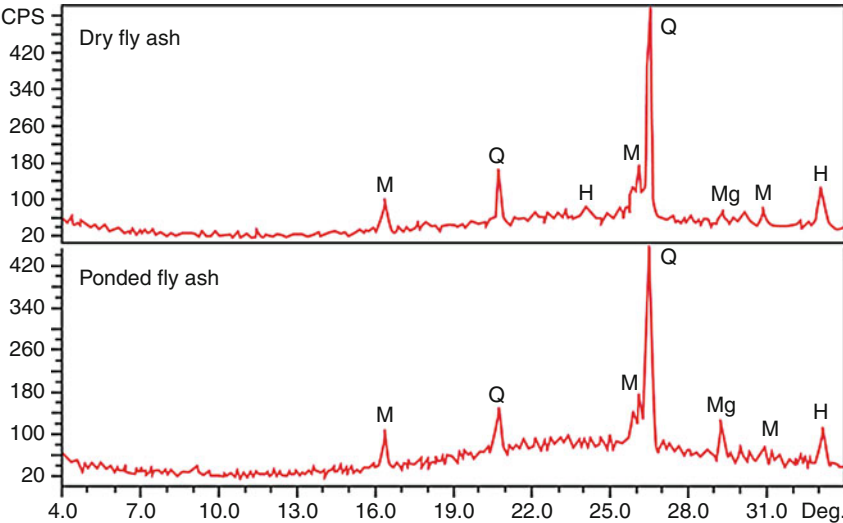
Compound	Bituminous ^a	Subbituminous ^b	Lignite ^c	ASTM C 618 specifications [18]	
	Class F	Class C	Class C	Class F	Class C
Silica (SiO ₂)	57.0	34.6	28.9		
Alumina (Al ₂ O ₃)	25.3	17.8	11.1		
Ferric oxide (Fe ₂ O ₃)	5.3	6.1	11.1		
SiO ₂ + Al ₂ O ₃ + Fe ₂ O ₃	87.6	58.5	54.1	70.0 min	50.0 min
Lime (CaO)	2.0	27.2	18.9		
Magnesia (MgO)	1.2	5.8	6.0		
Sodium oxide (Na ₂ O)	0.9	2.3	8.0	1.5 max	1.5 max
Potassium oxide (K ₂ O)	3.1	0.5	0.9		
Sulfur trioxide (SO ₃)	0.2	2.8	6.8	5.0 max	5.0 max
Loss on ignition (LOI)	3.1	0.9	0.4	6.0 max	6.0 max
Moisture content (%)	NV	NV	NV	3.0 max	3.0 max

NV not available

^aIllinois Basin coal fly ash [19]

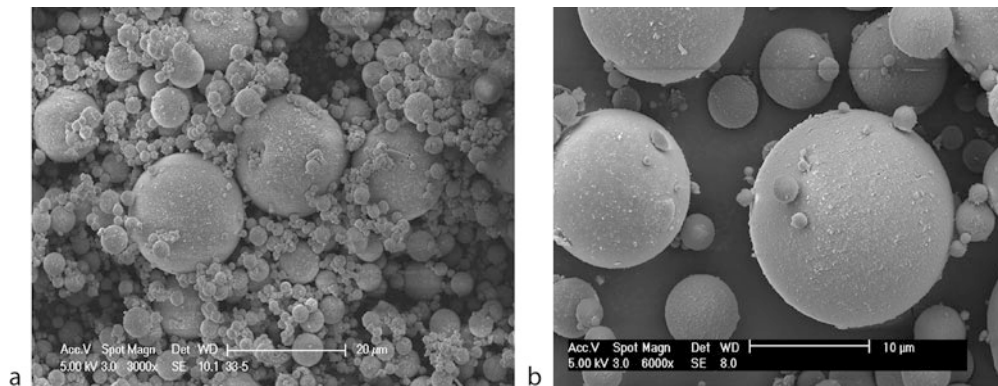
^bPowder River Basin coal fly ash [19]

^cNorth Dakota lignite coal fly ash [20]



Fly Ash. Figure 3

X-ray spectrogram of a dry fly ash sample and a ponded fly ash sample showing peaks for mullite (M), quartz (Q), hematite (H), and magnetite (Mg). The broad “hump” in the background between about 13° and 30° is due to abundant noncrystalline glass in the samples (printed with permission of the Illinois State Geological Survey) [17]



Fly Ash. Figure 4

SEM images of a bituminous coal fly ash sample at 3,000 $\times$ magnification (a) and 6,000 $\times$ magnification (b). Photos courtesy of the Illinois State Geological Survey

It should be noted that the LOI value correlates very well with the unburned carbon content, but it is not an actual measurement of the unburned carbon content in the sample. Burning lignite and subbituminous coals generally produces Class C fly ash that contains more calcium and magnesium and less unburned carbon (LOI) than Class F fly ash, which commonly results from burning bituminous and anthracite coal. Class F fly ash is typically pozzolanic in nature, but not cementitious by itself. Pozzolanic material contains glassy silica and alumina that in itself has little or no cementitious value, but in finely divided form and in the presence of water and free lime, will chemically react with the calcium hydroxide in the lime at ordinary temperatures to produce calcium aluminosilicate hydrates (cementitious compounds). On the other hand, typical Class C fly ash, in addition to having pozzolanic properties, has some cementitious properties, particularly the ability to increase concrete compressive strength in the presence of water alone, because of its higher lime (CaO) content. ASTM C 618 also sets the maximum acceptable LOI value (indicating unburned carbon content) at 6% for both Class F and Class C ash in concrete applications (Table 2) (see section “Concrete (Conventional) and Cement” for more discussion).

For use in Portland cement concrete, fly ash must meet the chemical (Table 2) and physical (Table 3) requirements specified by ASTM C 618. ASTM

specification C 311 provides standardized test methods for sampling and testing fly ash or natural pozzolans for use in Portland cement concrete [21].

Specifications have also been established in other countries or by industrial sectors [22–24]. In Canada, for example, specifications for the use of fly ash in concrete are given in CSA A3001 [25], which classifies fly ash further as Types F, CI, and CH based on the calcium oxide content. Type F refers to fly ash with a CaO content less than 8%, Type CI has a CaO content ranging from 8 to 20%, and Type CH fly ash has a CaO content greater than 20%. Higher CaO content generally indicates that the fly ash has higher cementitious properties. Two types of fly ash are produced and used in Canada, Type F and Type CI. Type CH fly ash is imported from the USA and used in concrete applications in Canada [26]. It is important to recognize that not all fly ashes meet the ASTM C 618 specifications. For applications other than concrete, fly ash may not need to meet ASTM C 618 requirements. Specifications of fly ash for other applications have also been developed, such as ASTM D 5239 for soil stabilization [27] and ASTM E 2277 for structural fill applications [28].

Applications

Fly ash, because of its mineralogical composition, fine particle size, and amorphous character, is commonly

Fly Ash. Table 3 Physical requirements of the ASTM C 618 specification [18]

Physical requirements	ASTM C 618 specifications	
	Class F	Class C
Amount retained when wet sieved on 45 μm sieve, max (%)	34	34
Pozzolanic activity index, with Portland cement at 28 days, min (%) of control	75	75
Pozzolanic activity index, with lime at 7 days, min (MPa)	5.5	–
Water requirement, max (%) of control	105	105
Autoclave expansion or contraction, max (%)	0.8	0.8
Specific gravity, max variation from average (%)	5	5
Percentage retained on 45 μm sieve, max variation,% points from average	5	5

pozzolanic and in some cases is also cementitious, whereas, bottom ash is much coarser and is not pozzolanic in nature. Therefore, it is important to distinguish fly ash from bottom ash. In addition, fly ash produced by different power plants commonly has different levels of pozzolanic and cementitious properties; therefore, fly ash must be analyzed and assigned for use in specific applications according to their analysis.

Fly ash has many established applications – mainly as an ingredient in concrete/cement products (see section “Concrete (Conventional) and Cement”), in stabilizing structural fills and embankments [29, 30], and in road bases and subbases [31]. Other uses are as a supplementary cementitious material in grout [32], soil stabilization [27, 33], and inclusion in flowable fill [34].

Fly ash also is used as mineral filler in asphalt concrete [35], roller compacted concrete [36–38], cured blocks, and fired brick products (sections “Autoclaved Aerated Concrete (AAC)” and “Ceramic Brick”), and also in the production of rubber [39, 40], paint [41, 42], and geopolymers [43–45]. It also has been found useful in agriculture [46–50] and in liquid and solid waste management [51, 52].

Another use for fly ash is for the extraction and recovery of high-value metals [53–58] that are minor or trace components of fly ash, such as magnetic iron (Fe), Aluminum (Al), Vanadium (Va), Strontium (Sr), and rare earth metals, such as Gallium (Ga) and Germanium (Ge).

Fly ash generated from some pulverized-coal-fired power plants contains particles of cenospheres

(gas-filled glassy spheres that float in water), which can be easily collected from the surface of ash ponds, or separated from dry fly ash using density separation device. Cenospheres have many unique properties including their light weight and hollowness, and almost perfectly spherical shape. They also have high strength compared to ordinary glass, and chemical resistance that enable them to be used in a wide range of commercial and industrial applications. They are especially useful as fillers in materials where their high strength, thermal insulating properties, and resistance to high temperatures are important. For more characteristics, advantages, and applications of cenospheres contact Fillite and Sphere Services, Inc. [59].

Fly ash applications by category and subcategory in various countries from most recent available sources [2, 7, 60–62] are tabulated in Table 4. (Information about such use categorization in China was not found.) The data indicate that the main use of fly ash falls into two categories: (1) concrete/cement products and (2) structural fill, embankments, and road base. In all cases, the use of fly ash as a component in concrete/cement products exceeds any other single application. The proportion of recycled fly ash used in concrete/cement products was 49.0% in India, 52.3% in the USA, 74.5% in the European Union countries, 85.4% in Canada, and 87.0% in Australia.

The governments of major fly ash-producing countries continue to promote use of coal fly ash and other CCPs through mandatory or incentive programs and research funding. Researchers continue to develop new and improved applications that allow fly ash to be used

Fly Ash. Table 4 Fly ash production and application by category and subcategory in various countries

Countries	India	USA	European Union	Canada	Australia
Year(s) and source	2004–2005 [7]	2008 [2]	2007 [60]	2004 [61]	2008 [62]
1 short ton = 0.9071 t (metric ton)	million tons	short tons	Kilo t	Kilo t	t
Total amount produced	112	100%	41,780	4,198	12,549,592
Total used in amount and in percentage	42	37.5%	20,015	1,300	2,018,854
Based on total amount used		100.0%	100.0%	100.0%	100.0%
Concrete/cement products	21	49.0%	15,766,509	1,110	1,757,379
Blended cement/cement material/clinker	N/A	N/A	8,281	638	N/A
Concrete/concrete products/grout	N/A	N/A	6,629	472	N/A
Structural fills/embankments, road base	18	42.0%	2,362	17	252,209 ^a
Structural fills/embankments	16	38.0%	2,071	–	N/A
Road base			291	17	N/A
Flowable fill	–	–	–	–	N/A
Dike	2	4.0%	–	–	N/A
General engineering fill	–	–	1,339	–	–
Aerated/non-aerated concrete blocks	–	–	1,151	–	–
Aerated concrete blocks	–	–	879	–	–
Non-aerated concrete blocks	–	–	272	–	–
Bricks + ceramics	1	2.0%	113	–	–
Waste stabilization/solidification	–	–	–	–	9,266
Mining applications	1	2.0%	–	89	0.5%
Miscellaneous/other	3	7.0%	140	84	6.5%

^aIncluding agriculture and aggregate applications

in greater quantities and in more valuable ways. A brief overview of the most often used and potentially high-volume established construction product applications of fly ash are described below.

Concrete (Conventional) and Cement

Historical background – Ash erupted from nearby volcanoes such as Mounts Etna and Vesuvius was used extensively in the construction of Roman structures. The Pantheon, which was built 2,000 years ago and is still standing today, is a classic example of durability achieved by using volcanic ash (see photo at <http://www.aaa-usa.org>). Fly ash, which was recognized as possessing pozzolanic properties similar to those of the volcanic ash that made ancient concrete structures durable, was used as an ingredient in concrete as early as 1914 [63]. However, the first comprehensive study of fly ash in concrete was described by R. E. Davis and others in 1937 in the USA [64]. Davis and his coworkers [64–66] studied the effects of the fineness, loss on ignition (LOI), and moisture content of fly ash on the weathering resistance and other properties of concrete and Portland cement. Their work provided a foundation that continues to be developed worldwide into modern specifications and quality assessments of fly ash-based products.

For example, the first British standard for the use of fly ash in concrete was established in 1965 and the first edition of standard BS 3892 [67] was as a fine aggregate (inert filler) and not as supplementary cementing material. During the time that research was underway, several comprehensive research projects were being conducted by the UK power industry to examine how base-load pulverized coal combustion conditions and fly ash collection systems influenced the physical and chemical properties of the fly ash produced and the results of those studies were also published in 1965 [68]. Standard BS 3892 was revised to include fly ash use as a cementitious component in structural concrete in 1982 [69]. Part I required fly ash to have a fineness limit of 12.0% by mass retained on a 45 μ m sieve and a LOI of less than 7.0%. This standard was further revised in 1997 with the same general property requirements, but the particle size range was expanded to allow up to 40% of the mass to be retained on a 45 μ m sieve [70]. The European fly ash for concrete

standard, BS EN 450, was established in 1994–1995 [71] and this standard was further revised into BS EN 450 -1 [72] and BS EN 450 -2 [73] to include definitions, specifications and conformity criteria, and evaluation, with some changes in fly ash categorization.

The construction of the Hungry Horse Dam in the USA completed in 1953, which used about 120,000 metric tons (132,240 short tons) of fly ash in the more than 3 million cubic yards of concrete to build dam, was the first major large-scale project that paved the way for using fly ash in the concrete industry [74, 75]. At least 100 other major locks and dams have been constructed using fly ash-containing concrete under the direction of the US Army Corps of Engineers, the US Bureau of Reclamation, or private engineering firms. Some of the most prestigious construction projects of recent times have relied on concrete containing fly ash, including dams, power stations, offshore platforms, tunnel, walls, highways, airport runways, commercial and residential buildings, bridges, pipelines, and silos. For worldwide examples and photos of some outstanding applications of fly ash in the construction industry see the ECOBA Web site [76] and the ACAA Web site [2].

Benefits – The unique benefit of using fly ash in concrete is that fly ash, like volcanic ash, is capable of reacting with calcium hydroxide (free lime) at ordinary temperature. When fly ash is used in combination with Portland cement, calcium hydroxide liberated from the hydration of Portland cement reacts with the aluminosilicates present in the fly ash to form calcium aluminosilicate hydrates (pozzolanic reaction products) possessing cementitious properties. This not only reduces the excess amount of calcium hydroxide (a nondurable by-product from hydration of Portland cement), but, in the pozzolanic reaction process, converts the calcium hydroxide into useful calcium aluminosilicate hydrates, which are the strongest and most durable portion of the paste in concrete. Also, pozzolanic reactions are much slower than cement hydration reaction thus the strength of concrete products is improving continuously.

Fly ash not only has been successfully used worldwide as a mineral admixture in Portland cement concrete for more than half of the past century, but also has been well established as a cement raw material and as a component of blended cement.

Fly ash used in concrete has been shown to have the following benefits: (1) decreases concrete's permeability, thus reducing the rate of penetration and erosion by water and resisting harmful alkali-aggregate and sulfate reactions; (2) increases the concrete mix's cohesiveness, thus reducing segregation and bleeding in the concrete; (3) contributes to long-term strength improvement in concrete; (4) improves the workability of fresh concrete, enabling more thorough compaction, and (5) reduces and delays the onset of heat of hydration in massive concrete structures such as dams and consequently the risks of thermal cracking of the concrete [77–80].

Overall, the greatest benefits in using fly ash in concrete/cement applications are environmental. Fly ash is itself a waste by-product material from coal-fired power plants. In the past, the impoundment and land disposal of raw fly ash has covered millions of acres of valuable land in the major fly ash-producing countries. Utilization of fly ash in valued products will reduce this landfill. Furthermore, when 1 ton of cement is manufactured, 0.85 t of CO₂ (a green house gas) are also produced and released into the atmosphere [81]. Using fly ash as a cement raw material or as a component of blended cement can reduce cement production, and thus reduce green house gas emissions.

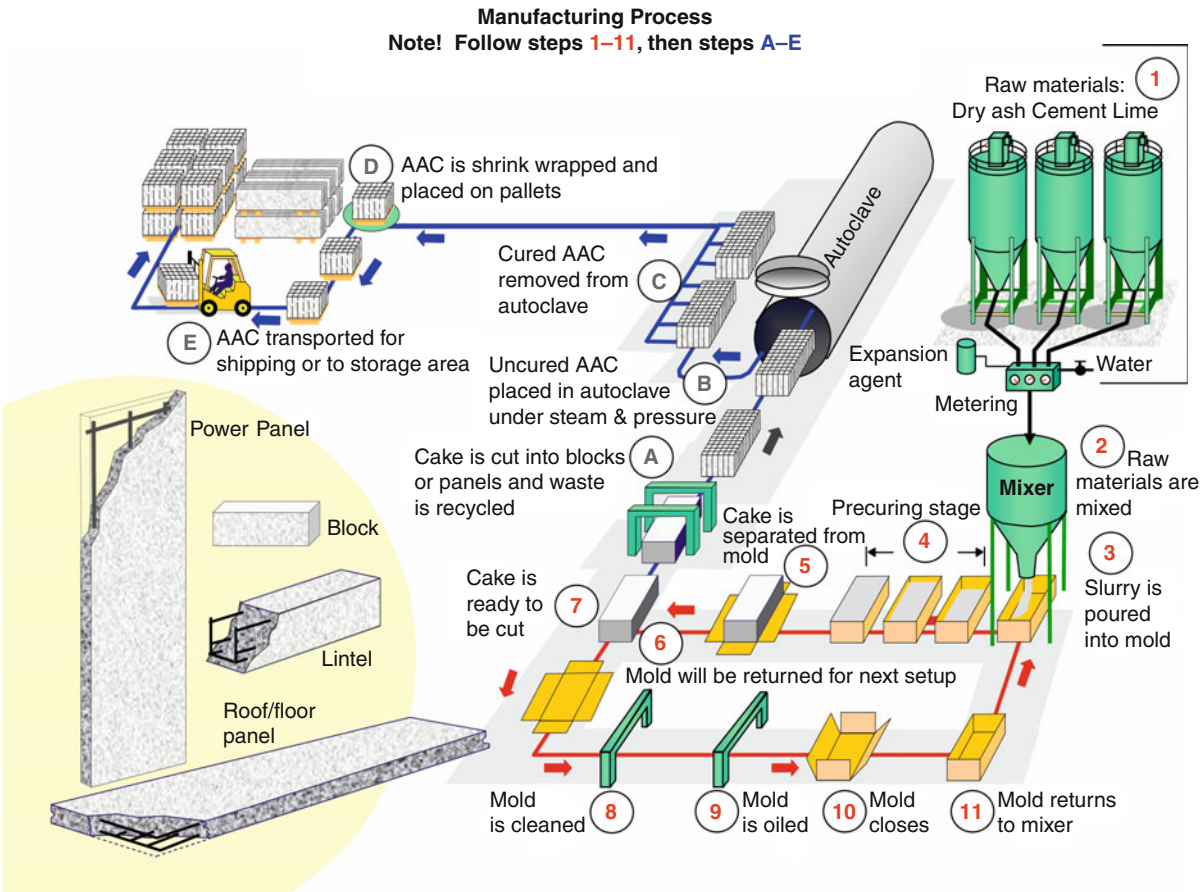
Fly ash properties in concrete applications – Fly ash properties may vary depending on the type of coal used and its chemical compositions and the burning conditions at a specific power plant. Class F fly ash, which contains low lime (CaO) content, is typically a result of burning anthracite or bituminous coal, whereas, high-lime Class C fly ash is generally from burning lignite or subbituminous coal. Fineness, loss on ignition (residue unburned carbon content), and chemical composition are the key elements of fly ash affecting its concrete applications. Fine fly ash generally gives resulting concrete high in compressive strength [82, 83]. In many concrete formulations, air bubbles are entrained into the mixture through mechanical agitation and the air bubbles are stabilized by an addition of specialized surfactants known as air-entraining admixtures (AEAs). The presence of small air bubbles increases the workability and cohesion of the concrete mix and improves the resistance of the set material to freeze–thaw cycles. The proper dosage of AEA is critical in

generating the desired amount of entrained air and stabilizing the air bubbles during mixing and transportation to the construction site [84]. Since the unburned carbon can adsorb the AEAs that are added to the mixture, the loss on ignition value of fly ash is restricted to no more than 6%. In terms of chemical composition, the fly ash used in concrete applications normally contains less than 10% of lime (CaO). Excessive lime in fly ash lowers the pozzolanic reactivity and affects the volume stability of the concrete. As mentioned in the section “[Fly Ash Properties and Classification](#),” fly ash used in Portland cement concrete applications must meet the chemical and physical requirements of ASTM C 618 or the specifications of other jurisdictions. Also, it must be in a dry form and its pozzolanic reactivity and quality consistency must be monitored closely. Those who want detailed information on fly ash for concrete and cement products can find references from Headwaters Resources [85], Turner-Fairbank Highway Research Center's coal fly ash user guide–Portland cement concrete [86], and from the book *Fly Ash in Concrete* by Gordon and Breach Science Publishers [87]. Other publications on the same subject can be obtained from the organizations listed in “Useful Resources” at the end of this entry.

Autoclaved Aerated Concrete (AAC)

Autoclaved aerated concrete (AAC), also known as autoclaved cellular concrete and porous concrete, is a lightweight building material that is made without any coarse aggregate. The manufacturing process of AAC is shown in [Fig. 5](#).

AAC is produced from a mixture of a silica-rich material (such as finely ground sand or fly ash as the major ingredient), lime and/or cement (as a binding agent), water, and very small amount of aluminum powder (as an expansion agent). The mixture is poured into large molds and allowed to expand as bread dough, forming a highly porous, but impermeable cellular structure (see [Fig. 6](#)). The cake (green aerated concrete) is separated from the molds and cut into standard size and shape for curing in a pressurized autoclave chamber at a temperature about 375°F. The ingredients combine during steam curing to form calcium silicate hydrates (the minerals typically formed from Portland cement), which give AAC its strength.



Fly Ash. Figure 5
Typical manufacturing process for producing autoclaved aerated concrete (AAC) (printed with permission of OSA Inc. [88])



Fly Ash. Figure 6
The highly porous structure of two AAC blocks produced at the Illinois State Geological Survey containing 64 wt% fly ash obtained from two different power plants (Photos courtesy of the Illinois State Geological Survey [90])

AAC has been produced in many countries for more than 70 years because it offers considerable advantages over other construction materials [89]. AAC is only one-fourth to one-third the weight of conventional concrete, has excellent thermal and sound insulation properties, is nonflammable, and insect resistant.

Sand-based AAC was developed during the 1920s (patented 1924), when Sweden was enduring an

extreme shortage of wood due to deforestation. In need of an alternative building material, Swedish architect Johan Eriksson developed AAC to meet the growing demand for a product that had wood's workability coupled with excellent thermal insulation, structural strength, and superior fire resistance. To further improve AAC, Hebel invented and patented a method of incorporating steel reinforcements during the manufacturing process in 1945 [89]. Since commercial production of the AAC material began in 1930, AAC has become one of the most commonly used building materials in Europe and its use is rapidly growing in many other countries. Fly ash can be used to substitute for the sand in the mixture, typically at 55–65% by dry weight. During the last decade, fly ash has become a widely used replacement for ground sand for making AAC in many countries.

In 1992, the UK used about 800,000 t of fly ash per year in the manufacture of approximately 1.5 million cubic meters of AAC, produced in ten factories [91]. In Czechoslovakia, six plants produced 2 million cubic meters of AAC, and 70% of this output was produced with fly ash [92]. The use of AAC has grown in recent years because of AAC's excellent thermal insulating properties and other economic and environmental advantages. In some countries, the demand for AAC products has exceeded its rate of production [93]. Currently, there are over 300 AAC manufacturing plants on six continents [94]. However, sand-based AAC production in the USA was not initiated until 1996 when ASTM commercial standards for AAC were established. There are currently three ASTM standards available for AAC – ASTM C1386 for material properties, ASTM C1452 for reinforced AAC panels, and ASTM C1555 for AAC masonry units [95].

In the 1990s, several studies were conducted in an attempt to establish fly ash-based AAC plants in the USA. A collaborative study led by the Electric Power Research Institute (EPRI) was begun in 1993. The project constructed an on-site AAC mobile pilot plant and tested various coal ashes for making AAC at nine utility plants of seven utility companies, including Tennessee Valley Authority (TVA) and Ohio Edison Co. [96]. Another study was initiated in 1997 by Babb International who built an AAC pilot plant in Ringgold, Georgia, [97] for the purpose of conducting test runs for making AAC using coal fly ash. Subsequently

during 1998–2000, Babb International built a commercial AAC plant in Indiana for producing fly ash-based AAC. However, both plants were closed in 2003, due in part to the slow acceptance for AAC in the USA [88].

From 1997 to 2005, six sand-based AAC plants were constructed in the USA but, as of 2009, only three of the six remain in limited operation serving a small AAC market. Extensive commercial use of AAC in the USA has not yet been achieved due in part to the low cost of lumber and the resistance of the home and commercial construction industry to switching raw materials and changing construction methods. However, this resistance may soon be overcome when architects, builders, and the public become more aware of the many benefits of using energy-efficient AAC materials. In 2005, OSA Inc. acquired the Ringgold AAC plant and, in collaboration with the Illinois State Geological Survey, developed an economical manufacturing method (2009 patent pending) for fly ash-based AAC production [98]. OSA Inc. is currently working with a local government in producing enough ash-based AAC materials for marketing analysis.

Ceramic Brick

The term “bricks” typically refers to ceramic (fired) bricks, the standard small building units in a rectangular shape, but the term has also been used to refer to unfired small building units in a rectangular shape, such as small concrete blocks, or (aerated or non-aerated) sand-lime bricks [99]. This section mainly considers the use of fly ash in making standard ceramic (fired) bricks.

In the process of manufacturing fired bricks, there are two critical steps – forming fired green bricks and firing them successfully. Clay and shale are the standard raw materials used in making conventional fired bricks. In general, green bricks are formed by two basic methods – mold-pressing (dry or wet) or extrusion (under vacuum) – and are then dried. Firing of the dried green bricks is done in a kiln or oven with proper temperature control for successful firing. In a typical firing cycle of progressively increasing temperature, the temperature rises fairly rapidly from ambient condition to a temperature of about 1,000°F and then at a slower rate up to about 1,500°F until the carbonaceous matter in clay/shale is oxidized completely [100].

Insufficient oxidation of the carbonaceous matter in the clay/shale mixture results in bricks produced with black or red cores. Also, improper rate of heating would result in defective bricks with cracks or bloating beneath the surface skin, which are commercially unacceptable. Iron oxide in the raw material is the primary cause of the typical redbrick color.

Concrete blocks are manufactured in a much wider range of shapes and sizes than clay bricks and are available with a wider range of treatments. A number of the treatments are used to provide a final size, color, and shape similar to that of fired clay bricks. These small concrete building blocks in a rectangular shape also are referred to as bricks. In this case, the “green” concrete bricks are formed in steel molds by vibration or by compaction with a hydraulic press. When dry, the green bricks rather than being fired are cured under low-pressure steam at ambient temperature to gain strength/hardness and other properties.

Fired brick manufacturing operations in most part of the world are labor- and energy-intensive. Brick manufacturers in developing countries, having cheaper labor, still rely heavily on mold-pressing of green bricks manually or by machine. In comparison, manufacturers in developed countries mainly utilize machine extrusion to form green bricks and tunnel kiln firing. A typical extruder machine produces about 8,000 green bricks per hour, which would take a team of two skilled brick makers 10 days to produce using mold-pressing method [101]. In 1998 in the USA, according to a survey conducted by the Brick Industry Association, 93.2% of total regular clay brick production was formed by extrusion. A machine-molded method was used for making specialty bricks that constituted 5.3% of the total production, and only 1.5% of all the manufactured bricks were produced by the dry-pressing method commonly used in developing countries [102]. Using the extrusion method, as much as 70 wt% of fly ash can be used to substitute for clay or shale; nearly 100% by weight can be substituted when using a mold-pressed method.

Fly ash and bottom ash were used in making ceramic (fired) products such as building bricks, pipes, and roof tiles as early as the 1930s, but the practices used and any valuable technical information gathered, in most cases, were not documented. These early trials generally took place as a result of informal

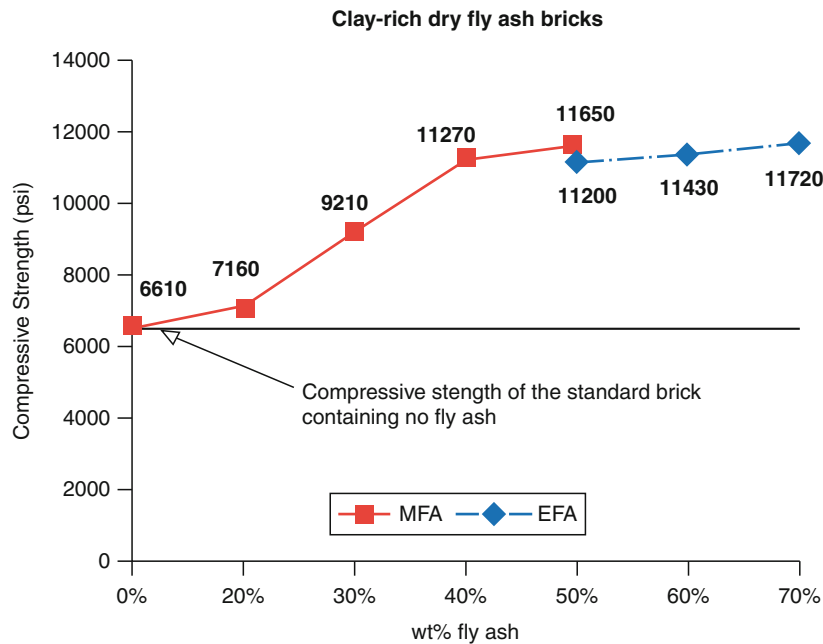
initiatives between a brick company and a local power plant and they had little incentive to preserve or publish results.

A brief historical review of making fired brick using fly ash was first conducted in the UK by Anderson and Jackson of the North Staffordshire Polytechnic in 1987 [103]. Their review indicated that despite successful trials and technical promotion during the late 1950s–1960s, the response by the brick industry to use fly ash in brick production was poor due in part to: (1) the high demand for fired bricks at that time, (2) the low cost of conventional raw materials, (3) the difficulties encountered in some field trials when making bricks containing fly ash, and (4) reservations about the consistency in the quality of the fly ash supply [103].

In 1977, Slonaker experimented with making fired bricks comprised of mainly of Class F fly ash generated from the West Virginia and Pennsylvania coals. He used a dry mold-pressed method to form green bricks, and acceptable bricks were produced from feeds consisting of as much as 72% fly ash, 25% bottom ash, and 3% sodium silicate [104]. However, Slonaker’s study did not lead to commercialization. His fly ash bricks did not have conventional redbrick color, and his bricks with a composition of nearly 100% ash probably lacked sufficient plasticity to be extruded well.

In 2005, Freight Pipeline Company (FPC) studied a mold-pressed method for making green bricks composed of almost 100% Class C fly ash. The brick strength, rather than firing, was developed by water-steam curing at room temperature [105]. Commercially available red-colored pigment was used as an additive to give the final brick products the traditional redbrick color. Recently, FPC issued its US patent license to CalStar Cement, a California-based company, for further development.

During 2000, researchers at the Illinois State Geological Survey, University of Illinois, began to experiment with using local Class F fly ash for manufacturing fired bricks. Fly ash was obtained as dry fly ash from hoppers or as ponded fly ash from a power plant pond or a stockpile. A systematic laboratory study on extruded fired bricks containing up to 70 wt% dry fly ash showed that the fired bricks had greater compressive strength (Fig. 7) [106] and lower thermal conductivity [107] than conventional bricks containing no fly ash.



Fly Ash. Figure 7

Compressive strength of fired bricks made with increasing amounts of fly ash in clay brick formulation. MFA, Plant-1 dry fly ash; EFA, Plant-2 dry fly ash. Figure courtesy of the Illinois State Geological Survey [106]

In further studies with industry partners, pilot-scale and commercial-scale tests, formulations, and procedures for producing commercial-size building and paving bricks containing amounts of fly ash were optimized to suit the plant's manufacturing processes. A handful of minor issues encountered in pilot-scale trials were resolved and commercial-scale production runs were successfully completed at three local brick plants [108–110]. Figure 8 shows a photo of an extrusion test run producing three-hole green fly ash building bricks at 8,000 bricks per hour (700 bricks per car). Figure 9 shows four batches of fired bricks successfully produced with fly ash inputs at 0% (E1), 20% (E2), 30% (E3), and 40% (E4) by volume. These studies, supported by state and federal governments, showed that Class F fly ash from Illinois Basin bituminous coals is suitable for fired brick production, especially when wet fly ash from retaining ponds was used. This is because most of the scum is associated with soluble salts in the dry fly ash and has been leached out and removed in the pond water. Thus, the amount of scum-suppressing additive otherwise required can be reduced or completely avoided.

Standard commercial-size fired bricks produced with clay/shale mixtures containing 30–40% by volume (28–37% by weight) of ponded fly ash showed color similar to conventional fired bricks and their engineering properties, including results of freeze–thaw tests, met or exceeded the ASTM C 62 commercial specifications [111]. Also, local builders acknowledged that their method for conventional brick construction can be applied directly to the construction of fly ash bricks.

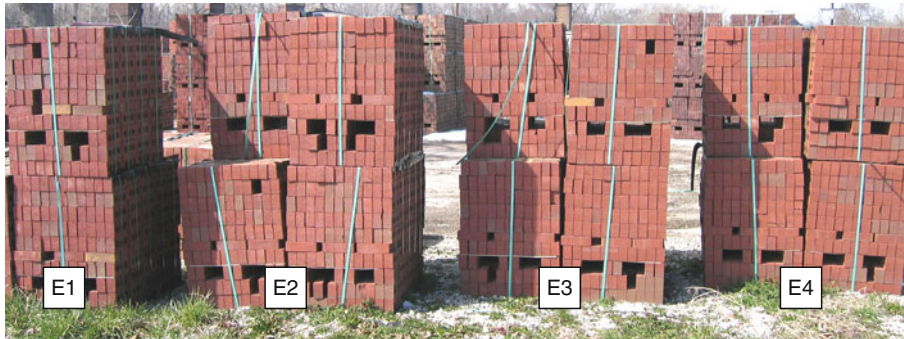
A knee-high wall at the Museum of Science and Industry in Chicago (Fig. 10) was constructed using Illinois fly ash bricks produced from commercial-scale production tests. This was part of the Museum's Clean Coal Technology Exhibition. The economic evaluation of the commercial-scale tests was optimistic [108]. An environmental assessment of fired bricks indicated that the fly ash-containing bricks were fully comparable to regular commercial bricks as environmentally safe construction materials [112].

Participating brick companies are seeking partners for a joint venture to commercially produce bricks containing fly ash. Only a few brick plants in the



Fly Ash. Figure 8

Scale-up extrusion demonstration at a brick plant producing three-hole green fly ash building bricks at 8,000 bricks per hour (700 bricks per car) (photo courtesy of the Illinois State Geological Survey [110])



Fly Ash. Figure 9

Four batches of fired building bricks produced from scale-up production test runs with fly ash inputs at 0% (E1), 20% (E2), 30% (E3), and 40% (E4) by volume (photo courtesy of the Illinois State Geological Survey [110])

USA are currently producing bricks containing small amounts (1–5%) of coal ash. In a sustainable world, this rate can be increased manyfold.

Environmental Impacts and Considerations

As mentioned earlier, use of fly ash as a recycled material can have economically and environmentally beneficial results. Use of fly ash as one of the ingredients in concrete to improve durability and service life is

important not only to private owners, but also to governmental authorities responsible for the design, construction, maintenance, and repair of buildings, bridges, roads, and infrastructure. The use of fly ash in concrete enhances the recycled material content of a building and is recognized as a beneficial strategy for CO₂ reduction. The use of fly ash in concrete-related products and bricks is an effective and increasingly used environmentally responsible strategy to promote sustainability because it (1) eliminates the need to



Fly Ash. Figure 10

A knee-high wall at the Chicago's Museum of Science and Industry constructed with fly ash bricks produced from commercial-scale production tests. This was part of the museum's Clean Coal Technology Exhibition (photo courtesy of the Illinois State Geological Survey)

dispose of fly ash, a typically industrial waste material; (2) reduces the cement content of concrete, and thus the CO_2 generated from cement manufacturing; (3) reduces the amount of embodied energy in the concrete; and (4) conserves natural resources by replacing the materials that would otherwise have to be mined to make these products. Also, the trace elements in the fly ash are stabilized in the concrete products, reducing the chances for releasing them into the environment [113]. A review on various reports and research papers in the UK regarding the environmental impacts of using fly ash was conducted by Sear and coworkers [114] in 2003. Their conclusion was that coal fly ash is an environmentally harmless material that can be safely used in bound (such as concrete products) and unbound (such as land reclamation) applications.

Potential health issues of using coal ash were also discussed in the Environmental Focus Report issued by the Electric Power Research Institute in 1998. The subjects included ash dust issue, leaching of coal ash from land applications and from products containing fly ash as well as skin contact and radioactivity of coal ash. It concluded that when handled and disposed properly, coal ash does not present a public health or environmental threat [115].

Having advantages in physical, chemical, and mineralogical properties, many fly ashes are marketed for

a wide range of engineering applications in construction and other industries. However, large volume of these materials is still remaining for disposal every year in the USA and the other major fly ash-producing countries. The environmental disposal practices of these materials are important issues impacting producers and regulators.

Physically, coal fly ash is a fine powdery material, and in its dry form, just like road dust, it readily becomes airborne even in a mild wind. Initially, coal-fired power plants released their fly ash together with flue gas into the atmosphere. The settlement of noticeable amounts of fly ash dust particles on the land had disturbed ecological systems and degraded agricultural land and soil quality. Health risks for utility workers from long-term exposure to fly ash dust, just like road dust, include diseases such as silicosis, fibrosis of the lungs, and bronchitis. Power plants eventually were mandated to use collection devices such as electrostatic precipitation systems to gather fly ash and keep it from being carried with exhaust gases out the stack. The general public does not now encounter coal ash in quantities that might result in health risks.

During power plant operation, some plant workers may be exposed to appreciable amounts of fly ash, either as suspended particulates or in collected ash piles. Such workers include primarily ash haulers, ash silo operators, truck loaders and drivers, ash system inspectors, and ash collection system personnel. The main sources of ash dust are the ash transfer system and ash loading. Under these circumstances, the precaution measures issued by the health organizations similar to the US Occupational Safety and Health Administration must be followed. However, studies on potential health effects of workers in frequent contact with coal fly ash showed that routine operating activities did not produce hazardous exposures, and occupational health records for these types of workers do not show a higher incidence of respiratory problems than those of power plant workers who do not work as closely with coal ash [115].

Dust control measures are also required at fly ash disposal sites to prevent dust hazard conditions and health risks. For example, transportation of fly ash to the disposal site typically is done in a conditioned form (moisture content about 15–30%) and in covered trucks or rail cars. At the disposal site, as the fly ash dries it forms a crust or clumps and is less easily moved

by the wind. In many cases, soil is applied on top of the fly ash as a cover material. These processes are designed to keep fly ash dust at a low enough level to avoid any expected health impacts.

Chemically, the main components of coal fly ash as indicated in Table 2 are about 60–90% silica, alumina, and iron oxides. Other metals and metal-like elements commonly are present only in trace quantities – arsenic, cadmium, beryllium, thallium, nickel, lead, manganese, chromium, selenium, zinc, and other metals. The primary health concern from the land disposal of fly ash is the potential for the trace elements listed above to enter nearby groundwater. Standardized leaching tests show that small amounts of toxic trace elements (particularly arsenic, selenium, lead, and mercury) may leach from fly ash that is in prolonged contact with ground water, but they do not migrate far from the ash site and are present in small enough concentrations to justify the regulatory classification of fly ash as nonhazardous solid waste [115]. Studies on various animals examining the toxic effects of ingesting fly ash constituents have not found pathological damage that would indicate potential human health problems [115]. The majority of fly ash is not enriched in radioactive elements, nor is it more radioactive, compared to common soils or rocks. Also, radiation in fly ash and in products produced from fly ash is minimal – well below the US Environmental Protection Agency's action standards [115, 116]. However, each fly ash should be analyzed on a case-by-case basis for potential toxins and radioactivity because coals come from many areas. Also, precaution measures should be undertaken when it is required by local governments.

Fly ash is not currently regulated in the USA by the US Environmental Protection Agency (EPA) nor in the other countries as a hazardous waste material; however, it must be handled or disposed of properly to prevent air pollution (irritant dusts) and groundwater contamination. For example, it is important to periodically monitor drinking water wells near fly ash disposal sites to ensure that the concentrations of trace elements remain below the regulatory levels for safe long-term drinking and cooking.

Another precautionary example for ensuring proper disposal of fly ash is evident by the failure of a retaining pond structure on December 22, 2008 at the Kingston power plant of the TVA in the USA. This structural failure

led to the release of approximately 5.4 million cubic yards of impounded fly ash onto surrounding land and into the adjacent Emory River. The cost for clean-up is estimated at more than \$1 billion, and costs of legal actions and settlement of possible lawsuits may add to this. This incident directly affects citizens/animals living in close proximity to the plant and the TVA, but also indirectly impacts all coal burning utilities and other large coal users throughout the USA. As a result of this incident, the US EPA instructed all power plant operators and companies to conduct on-site assessments to determine the structural integrity of their ash storage ponds, and take whatever measures were needed to ensure the integrity. Eventually, the US EPA may issue more rigorous regulations or may eliminate ash pond disposal altogether [117].

Future Directions

Although the use of nonfossil energy sources such as nuclear or wind is growing, coal will continue to be the major fuel for worldwide power generation for some time to come [118]. While significant progress has been made in terms of reusing fly ash combustion residues in the world, maintaining the quality consistency of a fly ash supply, which is extremely important in many already established applications, would require new strategies. This is because the utility industry starts using new improved coal combustion technologies and using more renewable energy sources, and endures more stringent emissions control regulations. Major developments that would alter fly ash properties and impact the use of fly ash in manufacturing products include, but are not limited to the following.

- The use of mercury reduction and/or NO_x reduction technologies (activated-carbon mercury adsorption, low NO_x burners, selective catalytic NO_x reduction, or selective non-catalytic NO_x reduction), would produce fly ash with unacceptably high levels of unburned carbon [119] and, in some cases, adsorbed mercury or ammonia.
- Switching coal sources and/or changing boiler operating conditions in an effort to reduce the cost for electricity generation would result in fly ash with inconsistent properties.
- Co-firing coal with other fuels (such as natural gas, wood, or other biomass) or other combustible wastes (such as petroleum coke, scrap tires,

municipal solid waste) would produce fly ash with inconsistent properties.

- The use of advanced clean coal technologies, such as gasification of coal or burning coal/petroleum coke mixed material, would produce residues, mainly a hard glassy slag, as well as some fly ash with compositions much different from conventional pulverized coal combustion fly ash.

Future developments for fly ash applications will require improved fly ash management strategies that will ensure the quality consistency of fly ash for new and established applications. These may include (1) blending and mixing of ashes for quality consistency, (2) particle size control by classification, grinding, and agglomeration, (3) physical or chemical processing for removal of residual carbon or ammonia, and (4) separation processing, including flotation and electrostatic or magnetic separation [120]. Some fly ash marketing companies have technologies ready for fly ash processing. But when additional treatment is not practical from an economic standpoint, the coal-fired power plant will have no other option other than land-fill disposal of the fly ash they produce. Managing fly ash in an efficient, environmentally sound, and cost-effective manner will remain a continuing challenge for the major fly ash producers, marketers, and regulators worldwide.

Bibliography

Primary Literature

- Joshi RC, Lohita RP (1997) Introduction, p1-4 in Fly ash in concrete: production, properties and uses. In: Malhotra VM (ed) *Advances in concrete technology*, vol 2. Gordon and Breach Science, Amsterdam
- American Coal Ash Association (2009) 2008 Coal combustion product (CCP) production & use survey report. 15200 E. Grand Ave. Ste 3050, Aurora, CO 80014. <http://www.acaa-usa.org>; the data in CCP Production & Use Reports
- Chikkatur AP, Sagar AD (2007) Cleaner power in India: towards a clean-coal-technology roadmap. Discussion paper 2007–06, Energy technology innovation policy. Belfer Center for Science and International Affairs, Kennedy School of Government, Harvard University, Cambridge
- Sinha R, Dayal U (2000) Save soil, use flyash. IIT Kanpur, sponsored by Flyash Mission, Department of Science & Technology at New Delhi. <http://www.iitk.ac.in/infocell/soil/testing2.htm>
- The Gazette of India (1999) Ministry of Environment and Forests notification, Part II, Section 3, sub-section (ii) S.O.763 (E), New Delhi, 14 Sept 1999. [http://envfor.nic.in/legis/hsm/so763\(e\).pdf](http://envfor.nic.in/legis/hsm/so763(e).pdf)
- Ministry of Environment and Forests (2003) Notification S.O. 979 (E): amendment. New Delhi, 27 Aug 2003. [http://envfor.nic.in/legis/hsm/so979\(e\).pdf](http://envfor.nic.in/legis/hsm/so979(e).pdf)
- Kumar V, Mathur M, Sinha SS, Dhatrak S (2005) Fly ash: an environment saviour. Fly Ash Utilization Programme (FAUP), TIFAC, DST, New Delhi, India. http://c-farm.org/001_FLY_ASH_AN_ENVIRONMENT_SAVIOUR.pdf
- (a) Maiti SK, Jaiswal S (2008) Bioaccumulation and translocation of metals in the natural vegetation growing on fly ash lagoons: a field study from Santaldih thermal power plant, West Bengal, India. *Environ Monit Assess* 136(1–3):355–370; (b) D'Monte D (2005) Some concrete solutions – recycling fly ash. *India Together – the news in proportion*, May 4, 2005
- China Building Material Federation (2010) Introduction on current fly ash research in China. http://www.coalash.org/content/cha_view.asp?id=495&cha_id=38
- Dhadse S, Kumari P, Bhagia LJ (2008) Fly ash characterization, utilization and government initiatives in India – a review. *J Sci Ind Res* 67:11–18
- Vejahati F, Xu Z, Gupta R (2010) Trace elements in coal: associations with coal and minerals and their behavior during coal utilization – a review. *Fuel* 89:904–911
- Chou MIM (2000–2008) Communications between Chou MIM and the managers of the utility plants in Central and Midwestern USA
- (a) DC industrial plant services private limited. <http://www.dcips.com/technology.html>; (b) FLSmidth Inc. (2005) Pneumatic transport for the fly ash management industry. <http://www.flsmidth.com/NR/rdonlyres/610A8705-A482-45E0-9DB1-633F2E0F597C/23644/AshHandling.pdf>
- Charah Inc. (2010) Ash management. <http://www.charah.com/AshManagementServices/AshManagement/tabid/76/Default.aspx>
- Turner-Fairbank Highway Research Center (TFHRC) (2010) Coal fly ash – material description. <http://www.tfhrc.gov/hnr20/recycle/waste/cfa51.htm>. TFHRC, 6300 Georgetown Pike, McLean, VA 22101, USA
- Manz O, Pflughoeft-Hassett D (2005) Historical perspective of coal ash marketing and promotion in the USA. In: *Proceeding: world of coal ash symposium*, Lexington, 11–15 Apr 2005. <http://www.flyash.info/2005/11man.pdf>
- Chou MIM, Chou SFJ, Chen LM, Stucki JM (2009) Manufacturing bricks with fly ash and advanced coal combustion by-products. Institute of Natural Resource Sustainability, Illinois State Geological Survey, Circular 574, 10 pp
- ASTM C 618 (2008) Standard specification for fly ash and raw or calcined natural pozzolan for use as mineral admixture in Portland cement concrete. American Society for Testing and Materials, West Conshohocken
- Unpublished data of Chou MIM. Illinois State Geological Survey
- The Fly Ash Resource Center – chemistry. <http://www.rmajko.com/chemistry.html>

21. ASTM C 311 (2007) Standard test methods for sampling and testing fly ash or natural pozzolans for use in Portland cement concrete. American Society for Testing and Materials, West Conshohocken
22. The Fly Ash Resource Center – specification. <http://www.rmajko.com/specs.html>
23. Recycled Materials Resource Center (2008) Evaluation guidance – standards and specifications. User guidelines for byproducts and secondary use materials in pavement construction, RMRC. <http://www.recycledmaterials.org/tools/uguidelines/standards.asp>
24. Wang L, Cui Y (2007) The application and development of fly ash in China. Presented at 2007 World Coal Ash (WOCA) conference, Covington, 7–10 May 2007
25. CSA A3001 (2004) Cementitious materials, New Canadian Standard: A3000-03 by Canadian Standards Association
26. Bouzoubaâ N, Foo S (2005) Use of fly ash and slag in concrete: a best practice guide. Final Report – MTL 2004-16 (TR-R) to the Materials Technology Laboratory, Program of Government of Canada Action Plan 2000 on climate change. <http://scm-ac.gc.ca/docs/bestpractices.pdf>
27. ASTM D 5239 (2004) Standard practice for characterizing fly ash for use in soil stabilization. American Society for Testing and Materials, West Conshohocken
28. ASTM E 2277 (2003) Standard guide for design and construction of coal ash structural fills. American Society for Testing and Materials, West Conshohocken
29. Turner-Fairbank Highway Research Center's coal fly ash user guide on embankment or fill and its references. <http://www.fhwa.dot.gov/publications/research/infrastructure/structures/97148/app4.cfm>
30. Federal Highway Administration (2006) Fly ash in structural fills/embankments. In: Fly ash facts for highway engineers. U.S. Department of Transportation, Washington, DC (Chap 6). <http://www.fhwa.dot.gov/pavement/recycling/fach06.cfm>
31. Turner-Fairbank Highway Research Center's coal fly ash user guide – stabilized base and its references. <http://www.fhwa.dot.gov/publications/research/infrastructure/structures/97148/cfa55.cfm>
32. Dodd WE (2005) Fly ash use in pressurized grout remote backfilling of abandoned underground mines in North Dakota. In: Interstate technical group on abandoned underground mines third biennial workshop. U.S. Department of Transportation, Federal Highway Administration, Kansas City, 25–27 Apr 2000. <http://www.fhwa.dot.gov/engineering/geotech/hazards/mine/workshops/kdot/kansas04.cfm>
33. Federal Highway Administration (2003) Fly ash in soil improvement. In: Fly ash facts for highway engineers. U.S. Department of Transportation, Washington, DC (Chap 7). <http://www.fhwa.dot.gov/pavement/recycling/fach07.cfm>
34. Turner-Fairbank Highway Research Center's coal fly ash user guide – flowable fill and its references. <http://www.fhwa.dot.gov/publications/research/infrastructure/structures/97148/cfa56.cfm>
35. Turner-Fairbank Highway Research Center's coal fly ash user guide – asphalt concrete and its references. <http://www.fhwa.dot.gov/publications/research/infrastructure/structures/97148/cfa52.cfm>
36. Golden DM, DiGioia AM (2003) Fly ash for highway construction and site development. Environmental Protection Agency (EPA), Coal Combustion Productions Partnership (C2P2), Coal Ash Utilization Case Study 10. <http://www.epa.gov/osw/partnerships/c2p2/cases/10-highway2.pdf>
37. Nagayama I, Jikan S (2003) 30 years history of roller-compacted concrete dam in Japan. In: Proceedings of the fourth International symposium on Roller Compacted Concrete (RCC) Dam, Madrid, 17–19 Nov 2003. <http://www.pwri.go.jp/eng/activity/pdf/reports/nagayama031116.pdf>
38. Abdo FY (Nov 2008) Roller-compacted-concrete dams: design and construction trends. Hydro Review, HCI Publications. <http://www.cement.org/water/RCC%20Design&Construction.pdf>
39. Saowapark T, Sombatsomporn N, Sirisinha C (2009) Viscoelastic properties of fly ash-filled natural rubber compounds: Effect of fly ash loading. J Appl Polym Sci 112(4):2552–2558
40. Kruger RA, Hovy M, Wardle D (1999) The use of fly ash fillers in rubber. In: Proceedings of 1999 international ash utilization symposium, Lexington, 18–20 Oct 1999. <http://www.flyash.info/1999/newprod/kruger2.pdf>
41. Pruett-Schaffer Paint Company. <http://www.undeerc.org/carr/buyersguide/product.asp?prodid=25&catid=33>
42. Yamazaki S, Okuda M (2006) Ship bottom paint using coal ash and diatomaceous earth. US Patent 7045788 B2, 23 May 2006
43. Rangan BJ (2008) Fly ash-based geopolymer concrete. In: Sustainable resource use. Your Building. <http://www.yourbuilding.org/article/NewsDetail.aspx?p=83&id=1570>
44. Louisiana Tech University (2009) Green research results in new geopolymer concrete technology. Science News, ScienceDaily. <http://www.sciencedaily.com/releases/2009/09/090929141534.htm>
45. Drechsler M (2006) Geopolymers: a new technology for sustainability in mining, construction and hazardous waste industries. Parsons Brinckerhoff PB Netw 62:72–74. http://www.pbworld.com/news_events/publications/network/issue_62/pb%20network%20issue%2062%20-%202025%20-%20geopolymers%20a%20new%20technology%20-%20drechsler.pdf
46. Beaver T (1995) Adding coal ash to the compositing mix. BioCycle 36(3):88–89
47. Renolds K, Kruger R, Rethman N (1999) The manufacture and evaluation of artificial soil (SLASH) prepared from fly ash and sewage sludge. In: 1999 international ash utilization symposium, Paper #1, Center for Applied energy research, University of Kentucky, Lexington, 18–20 Oct 1999
48. Mitra BN, Karmaker S, Swain DK, Gosh BC (2003) Fly ash – a potential source of soil amendment and a component of integrated plant nutrient supply. In: 2003 international ash utilization symposium, Center for Applied Energy Research, University of Kentucky, Lexington

49. Smith I (2005) Land uses of coal fly ash – benefits and barriers. International Energy Agency (IEA) Clean Coal Centre Report, CCC/96, London, p 30. ISBN 92-9029-411-6
50. Yunusa IAM, Eamus D, DeSilva DL, Murray BR, Burchett MD, Skilbeck GC, Heidrich C (2006) Fly-ash: an exploitable resource for management of Australian agricultural soils. *Fuel* 85: 2337–2344
51. Wang S, Viraraghavan T (1997) Wastewater sludge conditioning by fly ash. *Waste Manage* 17(7):443–450
52. Luna Y, Fernández-Pereira C, Vale JF, Alberca L (2009) Waste stabilization/solidification (S/S) using fly ash-based geopolymers. Influence of carbonation on the S/S of an EAF dust. Presented at 2009 World of Coal Ash (WOCA) conference, Lexington, 4–7 May 2009. <http://www.flyash.info/2009/097-fernandez2009.pdf>
53. Calzonetti FJ, Elmes GA (1981) Metal recovery from power plant ash: an ecological approach to coal. *GeoJournal* 3(1):59–70
54. McDowell WJ, Seeley FG (1981) Recovery of aluminum and other metal values from fly ash. US Patent 4252777, 24 Feb 1981
55. Blander M, Wai CM, Nagy Z (1984) Extraction of trace metals from fly ash. US Patent 4475993, 9 Oct 1984
56. Lisowyj B, Hitchcock DC, Epstein H (1987) Process for the recovery of gallium and germanium from coal fly ash. US Patent 4678647, 07 July 1987
57. Tsuboi I, Kasai S, Kunugita E, Komasaawa I (1991) Recovery of gallium and vanadium from coal fly ash. *J Chem Eng Jpn* 24(1):15–20
58. Cao DZ, Selic E, Herbell JD (2008) Utilization of fly ash from coal-fired power plants in China. *J Zhejiang Univ Sci A* 9(5):681–687
59. Cenospheres services companies: the Fillite <http://www.fillite.com/> and the Sphere Services, Inc. <http://www.sphereservices.com/>
60. European Coal Combustion Products Association (ECOBA) (2009) Production and utilization of CCPs in 2007 in Europe (E15). http://www.ecoba.com/evjm/media/statistics/ECOBA_Stat_2007_EU15.pdf
61. Panagapko D (2006) Cement. Canadian minerals year book 2006. <http://www.nrcan.gc.ca/mms-smm/busi-indu/cmy-amc/content/2006/17.pdf>
62. The Ash Development Association of Australia (ADAA) (2008) Annual membership survey results. http://www.adaa.asn.au/docs/ADAA_Mship_Report_2008.pdf
63. Anonymous (1914) An investigation of the pozzolanic nature of coal ashes. *Eng News* 71(24):1334–1335
64. Davis RE, Carlston RW, Kelly JW, Davis HE (1937) Properties of cements and concretes containing fly ash. *ACI J* 33:577–612
65. Davis RE, Kelly JW, Troxell GE, Davis HE (1935) Proportions of mortars and concretes containing Portland – pozzolan cements. *ACI J* 33:577–612
66. Davis RE, Davis HE, Kelly JW (1941) Weathering resistance of concretes containing fly-ash cements. *ACI J* 37:281–296
67. BSI, BS 3892 (1965) Pulverized fuel ash for use in concrete. British Standards Institution, London
68. Sear LKA (2001) UK practice – a review of fly ashes for use in concrete. Proceedings of workshop on novel products from combustion residues, Morella, Spain, 6–8 June 2001
69. BSI, BS 3892 (1982) Part 1: pulverized fuel ash for use as a cementitious component in structural concrete. British Standards Institution, London
70. BSI, BS 3892 (1997) Part 1: specification for pulverized fuel ash for use with Portland cement. British Standards Institution, London
71. BSI, BS EN 450 (1995) Fly ash for concrete – definition, requirements, and quality control. British Standards Institution, London
72. BSI, BS EN 450-1 (2005) Fly ash for concrete – definition, requirements, and conformity criteria. British Standards Institution, London
73. BSI, BS EN 450-2 (2005) Fly ash for concrete – conformity evaluation. British Standards Institution, London
74. FHWA. Infrastructure – materials group, fly ash. Federal Highway Administration, Department of Transportation. <http://www.fhwa.dot.gov/infrastructure/materialsgrp/flyash.htm>
75. Sutton I (1968) Geographical aspects of construction planning: Hoover Dam revisited. *J West* 7(3):301–344
76. ECOBA website. <http://www.ecoba.com/ecobaccpexs.html>
77. Taylor HFW (1997) Cement chemistry, 2nd edn. Thomas Telford, London
78. Papadakis VG (1999) Effect of fly ash on Portland cement systems, Part I. Low-calcium fly ash. *Cem Concr Res* 29(11):1727–1736
79. Papadakis VG (2000) Effect of fly ash on Portland cement systems, Part II. High-calcium fly ash. *Cem Concr Res* 30(10):1647–1654
80. Build It Green (2005) Fly ash concrete. <http://www.builditgreen.org/attachments/wysiwyg/22/Fly-Ash-Concrete.pdf>
81. Van Oss HG, Padovani AC (2003) Cement manufacture and the environment, Part II: environmental challenges and opportunities. *J Ind Ecol* 7(1):93–126
82. Erdoğan K, Türker P (1998) Effects of fly ash particle size on strength of Portland cement fly ash mortars. *Cem Concr Res* 28(9):1217–1222
83. Chindapasirt P, Jaturapitakkul C, Sinsiri T (2005) Effect of fly ash fineness on compressive strength and pore size of blended cement paste. *Cement Concr Compos* 27(4): 425–428
84. Gao YM, Shim HS, Hurt RH, Suuberg EM (1997) Effects of carbon on air entrainment in fly ash concrete: the role of soot and carbon black. *Energy Fuels* 11(2):457–462
85. Headwater Resources. Fly ash for concrete. http://www.flyash.com/data/upimages/press/HWR_brochure_flyash.pdf
86. Turner-Fairbank Highway Research Center's coal fly ash user guide – Portland cement concrete and its references. <http://www.fhwa.dot.gov/publications/research/infrastructure/structures/97148/cfa52.cfm>

87. Joshi RC, Lohita RP (1997) Fly ash in concrete: production, properties and uses. In: Malhotra VM (ed) *Advances in concrete technology*, vol 2. Gordon and Breach Science, Amsterdam, pp 269
88. OSA Inc. (2009) Autoclaved aerated concrete, oral presentation at the Tennessee Valley Authority, June 2009
89. Autoclaved Aerated Concrete Products Association (AACPA). AERCON AAC technical manual, introduction. <http://www.aacpa.org/>; <http://www.aerconaac.com/TECHNICAL%20MANUAL/Introduction.pdf>
90. Chou SFJ, Chou MIM, Bukowski JM (2003) Manufacturing energy efficient construction products from Illinois coal fly ash, Final technical report to the Illinois Clean Coal Institute. Illinois State Geological Survey
91. Carroll RA, Payne JC (1992) Autoclaved cementitious products using pulverized fuel ash. In: Wittmann FH (ed) *Advances in autoclaved aerated concrete*. AA Balkema, Rotterdam, pp 297–303
92. Hums D (1992) Ecological aspects for the production and used of autoclaved aerated concrete. In: Wittmann FH (ed) *Advances in autoclaved aerated concrete*. AA Balkema, Rotterdam, pp 43–49
93. George J (2008) Cement crisis doubles demand for AAC. Article on Emirates Business 24|7, 21 Aug 2008. http://www.business24-7.ae/Articles/2008/8/Pages/08212008_73fd4cfa6e6744ac80024a2f2a8f6bdd.aspx
94. Masonry Construction (2006) Autoclaved aerated concrete: product offers tremendous business opportunities for the masonry contractor. http://findarticles.com/p/articles/mi_m0NTA/is_3_19/ai_n26815025/
95. (a) ASTM C 1386 (2007) Standard specification for precast Autoclaved Aerated Concrete (AAC) wall construction units. American Society for Testing and Materials, West Conshohocken; (b) ASTM C 1693 (2009) Standard specification for Autoclaved Aerated Concrete (AAC). American Society for Testing and Materials, West Conshohocken; (c) ASTM C 1691 (2009) Standard specification for unreinforced Autoclaved Aerated Concrete (AAC) masonry units. American Society for Testing and Materials, West Conshohocken
96. Horvath ML (1995) Investigation of the use of fly-ash based autoclaved cellular concrete blocks in coal mines for air duct work, 25 Jan 1993–31 Dec 1994. Ohio Edison Company, Final report to the Ohio State Coal Development Office. <http://www.ohioairquality.org/ocdo/pdf/d92102.pdf>
97. Babb Cellular Concrete/Babb International (1998) News release. The Ringgold micro-plant, Ringgold
98. Chou MIM, Babb RA, Chou SFJ (2009). Concrete manufacturing processes and methods. United States Patent Application No. 12/573,475, University of Illinois at Urbana-Champaign and OSA Inc
99. Cicek T, Tanriverdi M (2007) Lime based steam autoclaved fly ash bricks. *Constr Build Mater* 21(6):1295–1300
100. Brick Industry Association (2006) Manufacturing of brick. Technical notes on brick construction TN 9, 9 Dec 2006. Brick Industry Association, Virginia
101. Regional Wood Energy Development Program in Asia (1993) Status and development issues of the brick industry in Asia, Report: GCP/RAS/154/NET: sponsored by the Food and Agriculture Organization of the United Nations, 69 pp
102. Brick Industry Association (1998) Manufacturing, classification, and selection of brick. Technical notes on brick construction TN 9, Manufacturing Part 1. Brick Industry Association, Virginia
103. Anderson M, Jackson G (1987) The history of pulverized fuel ash in brickmaking in Britain. *Trans J Br Ceram* 86(4):107–112
104. Slonaker JF (1977) The role of fly ash brick manufacturing in energy conservation. Report No. 149. Coal Research Bureau, College of Mineral and Energy Resources, West Virginia University, Morgantown, West Virginia, pp 6
105. Liu H (2005) Compacting fly ash to make bricks. Final technical report to National Science Foundation, Mar 2005, 15 pp
106. Chou MIM, Chou SFJ, Patel V, Botha F (2003) Production and testing of fired bricks made with Class F fly ash from two Illinois coal-fired power plants. Preprint and oral presentation at the 15th international symposium on management and use of coal combustion products held by the American Coal Ash Association, St. Petersburg, 27–30 Jan 2003
107. Chou MIM, Patel V, Laird CJ, Ho KK (2001) Chemical and engineering properties of fired bricks containing 50 weight percent of Class F fly ash. *Energy Sources* 23:665–673
108. Chou MIM, Chou SFJ, Patel V, Stucki JW, and Botha F (2003) Manufacturing commercial bricks with Illinois coal fly ash—a program overview. In: *International coal ash utilization symposium*, American Coal Ash Association, Lexington, 20–24 Oct 2003
109. Chou MIM, Chou SFJ, Patel V, Stucki JW (2005) Commercial production of fired bricks with Illinois Basin Class F fly ash. Preprint and oral presentation at the international congress on fly ash utilization conference, New Delhi, 4–7 Dec 2005
110. Chou MIM, Chou SFJ, Patel V, Pickering MD, Stucki JW (2006) Manufacturing building bricks with Class F fly ash from Illinois Basin coal, Final Report to US DOE/WVURC-combustion byproducts recycling consortium. Project Number: 02-CBRC-M12. Illinois State Geological Survey
111. ASTM C 62 (2008) Standard specification for building brick (solid masonry units made from clay or shale). American Society for Testing and Materials, West Conshohocken
112. Chou MIM, Chou SFJ, Pickering MD, Stucki JW (2006) An environmental feasibility assessment of fired bricks containing fly ash. *Am Chem Soc Div Fuel Chem* 51(2):421–422 (Preprint paper)
113. Zhang MH, Blanchette MC, Malhotra MV (2001) Leachability of trace metal elements from fly ash concrete: results from column-leaching and batch-leaching tests. *ACI Mater J* 98(2):126–136
114. Sear LKA, Weatherley AJ, Dawaon A (2003) The environmental impacts of using fly ash – the UK producers' perspective. In: *The 2003 international ash utilization symposium paper*

- #20, Center for Applied Energy Research, University of Kentucky, Lexington. Paper available at: <http://www.flyash.info>
115. Electric Power Research Institute (1998) Coal ash: its origin, disposal, use, and potential health issues. Electric Power Research Institute (EPRI) Environmental Focus Report, BR-111026. http://acaa.affiniscape.com/associations/8003/files/epri-coal_ash-origin-disp-use.pdf
 116. United States Geological Survey (1997) Radioactive elements in coal and fly ash: abundance, forms, and environmental significance. United States Geological Survey (USGS): USGS Fact Sheet FS-163-97. <http://pubs.usgs.gov/fs/1997/fs163-97/FS-163-97.pdf>
 117. Morford S (2009) EPA study finds dangers in coal ash ponds nationwide. Daily Climate News and Analysis, 1 Sept 2009. <http://solveclimate.com/blog/20090901/epa-study-finds-dangers-coal-ash-ponds-nationwide>
 118. Newell R (2009) U.S. Energy Information Administration – independent statistics and analysis. Annual energy outlook 2010 – reference case. <http://www.eia.gov>
 119. Pedersen KH, Jensen AD, Skjoth-Rasmussen MS, Dam-Johansen K (2008) A review of the interference of carbon containing fly ash with air entrainment in concrete. Prog Energy Combust Sci 34(2):135–154
 120. Hall ML, Livingston WR (2002) Fly ash quality, past, present, and future, and the effect of ash on the development of novel products. J Chem Technol Biotechnol 77:234–239
- Buyer's guide to coal ash containing products. <http://www.undeerc.org/carrc/buyersguide/default.asp>
- Coal Combustion Products Partnership (C²P²) – in the USA. <http://www.epa.gov/waste/partnerships/c2p2/index.htm>
- Electric Power Research Institute – in the USA. <http://www.epri.com>
- European Coal Combustion Products Association (ECOPA) – in Germany. <http://www.ecoba.com/>
- Fly Ash Information Center – in India. <http://www.fly-ash-information-center.org.in/>
- Fly Ash Library <http://www.flyash.info/> selected papers from the international ash utilization symposiums held in 1999, 2001, 2003 and world of coal ash conferences held in 2005 and 2007
- Fly Ash Utilization Programme (FAUP) – in India. <http://flyashindia.tifac.org.in/>
- Headwaters Resources – fly ash for concrete. http://www.flyash.com/data/upimages/press/HWR_brochure_flyash.pdf
- Informational & Analytical Center “Ecology of Power Engineering” of Moscow Power Engineering Institute (IACEE MPEI). <http://www.ecopower.ru/index.php?do=cat&category=english>
- Japan Coal Energy Center. <http://www.jcoal.or.jp/index-en.html>
- KEMA – a commercial service company, owned by several Dutch utilities, but functioning independently. <http://www.kema.com/services/consulting/hse/cc-consequences/fly-ash.aspx>
- Korean Coal Ash Recycling Association. http://coalash.or.kr/en/intro_3.html
- National Coal Ash Board, Israel. <http://www.coal-ash.co.il/english/index.html>
- RockTron – a UK company whose mission is to transform fresh and stockpiled fly ash from coal-fired power stations into a sustainable, alternative source of valuable, environmentally-coherent minerals for industrial use. <http://rktron.com/>
- South African Coal Ash Association (SACAA). <http://www.coalash.co.za/>
- The Coal Ash Research Center – at the University of North Dakota, Energy & Environmental Research Center (EERC) is dedicated to improving the technical and economic aspects of by-products management. [http://www.undeerc.org/carrc/](http://www.undeerc.org/carrc/Technology%20Information,%20Forecasting%20and%20Assessment%20Council%20(TIFAC))
- Technology Information, Forecasting and Assessment Council (TIFAC) <http://www.tifac.org.in> (established in 1988 in India)
- The Fly Ash Resource Center – in the USA whose mission is to promote knowledge and utilization of fly ash and other coal combustion products. <http://www.rmajko.com/flyash.html>
- United Kingdom Quality Ash Association (UKQAA). <http://www.ukqaa.org.uk/>
- United States Army Corps of Engineers. http://en.wikipedia.org/wiki/United_States_Army_Corps_of_Engineers
- United States Bureau of Reclamation. <http://www.usbr.gov/>
- United States Department of Transportation – Federal Highway Administration (FHWA): the Turner-Fairbank Highway Research Center. <http://www.tfhrc.gov>
- United States Geological Survey. <http://www.usgs.gov>
- World Wide Coal Combustion Product Network (WWCCPCN). <http://www.wwccpn.org/>

Books and Reviews

- Cox M, Nugteren H, Janssen-Jurkovičová M (2008) Combustion residues: current, novel and renewable applications. Wiley, England
- Manz OE (1997) Worldwide production of coal ash and utilization in concrete and other products. Fuel 76(8):691–696
- Sear LKA (2001) The properties and use of coal fly ash. Thomas Telford, London
- Taylor HFW (1997) Cement chemistry, 2nd edn. Thomas Telford, London
- Vassilev SV, Vassileva CG (2005) Methods for characterization of composition of fly ashes from coal-fired power stations: a critical overview. Energy Fuels 19(3):1084–1098
- Wesche K (1992) Fly ash in concrete: properties and performance. Rilem Report 7, the International Union of Testing and Research Laboratories for Materials and Structures. E & FN SPON, London

Useful Resources

- American Coal Ash Association (ACAA). <http://www.acaa.org>
- American Concrete Institute. <http://www.aci.org>
- Ash Development Association of Australia (ADAA). <http://www.adaa.asn.au/>
- Association of Canadian Industries Recycling Coal Ash (CIRCA). <http://circainfo.ca/>
- Boral Material Technologies Inc. (Boral MTI) – a US company. <http://www.boralmti.com> processing and marketing fly ash

Fossil Energy, Introduction

RIPUDAMAN MALHOTRA

Chemistry and Chemical Engineering Laboratory,
Pure and Applied Physical Sciences Division,
SRI International, Menlo Park, CA, USA

To many in the sustainability community, fossil energy is an anathema. Continued use of fossil resources – oil, coal, and natural gas – poses threats to the environment through the emission of pollutants and greenhouse gases. The fact that they are a limited or exhaustible resource means that in the future we could either run out of them or their extraction will get progressively harder to a point that it takes more energy to extract them than would be derived from their use. Using fossil energy is clearly not sustainable, and the world has to look to renewable resources for long-term survival.

This is an encyclopedia about the science and technology of sustainability. The word, *sustainability* shares its root with *sustenance*. Any discussion of sustainability must therefore include discussions of sustenance, and in the context of modern society, sustenance stems from the use of energy. We derive energy from a number of different sources. Annual global consumption of energy is currently on the order of 500 exajoules (EJ), about 85% of which comes from fossil resources. From 500 EJ/year, the global energy demand is expected to rise to somewhere between 1,000 and 1,500 EJ/year by the middle of this century. The drivers for this increased demand are already in place. Large segments of China, India, and Brazil are poised to increase their standard of living – and the concomitant energy demand – substantially. If the average energy consumption in the Asia-Pacific region were to reach the current global average, which is about half of what is consumed in Europe and Japan, the demand would increase by an amount equal to the total energy consumption in the North America region. The challenge of achieving a sustainable future is in being able to balance the energy requirements for the billions of people so they can lead healthy and productive lives against the need to preserve the environment by not running into its limits in its ability to supply the resources or act as a sink for the waste [1]. The entries in this section examine the current status, assess the

resource potential of the various fossil fuels, examine the technologies for using them, and review the environmental impact of their use.

Petroleum and natural gas have similar origin and often occur together in geologic formations, and their global distribution and production is discussed together in the entry by McCabe (► [Oil and Natural Gas: Global Resources](#)). This entry makes clear the distinction between reserves and resources. Reserves represent only that fraction of the resource base that can be economically recovered using current technology. These are not fixed quantities as both technology and economics change over time. Global annual production (and consumption) of oil in 2010 was 31 billion barrels and of natural gas was around 120 trillion cubic feet. In energy units, they correspond to 180 EJ of oil and 120 EJ of natural gas [2]. The current reserves are estimated at 1,236 billion barrels of conventional oil (7,500 EJ) and 6,545 tcf of natural gas (6,500 EJ). The current reserves to production ratio (R/P) is about 40 for oil and about 55 for natural gas. The R/P ratio has often been mistakenly taken as the time to exhaustion, but new discoveries as well as advances in technology add to the reserves. In the case of oil, for example, the R/P ratio has stayed around 40–50 years for more than 60 years even with the steadily increasing oil consumption. In addition, there are also unconventional accumulations of these hydrocarbon resources and extracting them requires development of new technologies. In the case of oil, the unconventional resources are oil sands, oil shale, and heavy oil. Unconventional resources of natural gas are coal bed methane, tight gas, shale gas, and gas hydrates. These unconventional resources are vast and have the potential of more than doubling our resource endowment.

Exploration and production of oil from sedimentary deposits and oil sands is the subject of an entry by Speight (► [Petroleum and Oil Sands Exploration and Production](#)). The processes for recovering oil could be a simple matter of drilling into the formation with the oil flowing to the surface under its own pressure, or it may require injection of gases, fluids, and surfactants to coax it to flow. In extreme cases, it may even require underground combustion of a portion of the resource to release the oil. Speight describes the chemical and physical factors that govern the flow of oil and the

technology options currently available. In a different entry, Speight (► [Petroleum Refining and Environmental Control and Environmental Effects](#)) provides and account of the different processes such as distillation, catalytic cracking, hydrotreating, reforming, and deasphalting used in the refining of crude oil. This entry also deals with environmental effects of the gaseous, liquid, and solid effluents from these processes. Production of oil from shale is principally achieved by retorting of shale, or other thermal processes including in situ pyrolysis. Oja and Suuberg (► [Oil Shale Processing, Chemistry and Technology](#)) detail the chemistry and technology of these processes in the entry.

Gas hydrates represent a particularly noteworthy resource and are reviewed in the entry by Patil (► [Alaska Gas Hydrate Research and Field Studies](#)). They are naturally occurring ice-like substances in which methane or other light gases are trapped in the cage structures formed by water molecules. They are stable under certain conditions of temperature and pressure, and can be found at many places around the world at depths of several hundred meters below the seabed and in the permafrost in the polar region. Global estimates of gas hydrates are on the order of 500,000 tcf, about a hundred times the proved reserves of natural gas. Apart from being a potentially very important source for energy in the future, the gas hydrates are also important from the perspective of climate change. A general warming could release substantial amounts of methane, which – being a potent greenhouse gas – would reinforce any global warming.

Oil remains the prized fuel. It has a high energy density and as a liquid it is well suited for transportation. The transportation sector relies on oil for over 90% of its energy needs (the remainder being mostly furnished by coal – via electricity) [3]. Given the importance of liquid fuels, there is considerable interest in converting coal and natural gas, the other hydrocarbon resources, into oil. In a pair of entries, Araso and Smith (► [Gas to Liquid Technologies](#) and ► [Coal to Liquids Technologies](#)) provide an overview of the various processes for converting natural gas and coal to liquids. These reviews cover the basic chemistry and catalytic technologies behind the different approaches. The entry on ► [Gas to Liquid Technologies](#) deals with steam reforming of methane, autothermal reforming, and partial oxidation approaches for producing syngas,

including strategies for managing the heat and mass transfer through the use of different reactor technologies. The entry next covers the conversion of syngas into liquids by Fischer-Tropsch (FT) synthesis. The entry on ► [Coal to Liquids Technologies](#) reviews the different approaches like pyrolysis, direct liquefaction, coprocessing with petroleum, and indirect liquefaction that first converts coal into syngas and then uses FT or other conversion processes to make liquid fuels.

The internal combustion engines that power these vehicles convert only about 20–30% of the energy in the fuel to motive power. Increasing the efficiency of the engines used in transportation represents a significant opportunity to reduce future demand for oil, as well as reduce the carbon footprint of the cars, trucks, and planes. Jacobs (► [Internal Combustion Engines, Developments in](#)) provides a detailed account of the thermodynamics of the different kinds of engines and of the various technologies under development for improving the efficiency of the internal combustion engines. These include strategies such as engine downsizing, turbocharging, better controls, variable geometry engines, variable valve timing, homogeneous charge compression engines, and waste heat recovery.

Höök (► [Coal and Peat: Global Resources and Future Supply](#)) reviews the global coal and peat resources, which are substantially larger than conventional oil and gas resources. Peat and coal are also distributed more evenly around the globe. The proved reserves of coal are estimated at between 800 billion and a trillion metric tons, representing roughly 20,000 EJ. Most coal is derived from trees and ferns that grew some 250–300 million years ago during the carboniferous period. Aerobic bacteria generally decompose most plant matter into CO₂, but when a tree fell into a swamp and was buried with limited exposure to oxygen, it could be partially preserved. That phenomenon occurs even today and is evidenced in the formation of peat (very young coal) in bogs. Unlike coal, peat has a low calorific content and its commercial use as an energy resource is limited. However, peat is important from global carbon cycle. Peat bogs store vast amounts of carbon, but the product of anaerobic decomposition, methane, contributes to greenhouse gases in the atmosphere such that under steady state conditions they comprise only a slight sink of carbon. The situation changes with

fires as peat bogs become major source of carbon in the atmosphere. Human activities, such as drainage of the area for agriculture, exacerbate the situation, as more peat gets exposed and the swamp is not wet-enough to squelch the fire. Estimates of the amounts of carbon released from peat are on the order of several billion tons, the same order of magnitude as from fuels used in transportation or power production.

As coal resources are vast, they are likely to continue to contribute substantially to global energy mix. However, their use also presents a number of challenges, beginning with mining, and through coal preparation and use. Mining operations result in excavation of billions of tons of earth, and depending on the nature of the waste rock and how it is handled it can lead to contamination of air, water, and soils systems. The wide-ranging nature of environmental impact necessitates a commensurately broad portfolio of technologies for the remediation, restoration and reclamation. Chattopadhyay and Chattopadhyay ([► Mining Industries and Their Sustainable Management](#)) provide an overview of the environmental impact of coal mining, and mining processes in general, as well as review the range of the chemical and biochemical strategies to mitigate the impact.

The run-of-mine coal is often laden with noncombustible minerals such as shale and clays that reduce its heating value. Cleaning and washing of coals removes many of these impurities and upgrades the coal into a marketable resource, improve the performance of power plants, and also reduce potential of harmful emissions and dust. Luttrell and Honaker ([► Coal Preparation](#)) review these cleaning and preparation operations, and point out the importance of cheap preparation in extending the resource base.

Electricity is one of the most useful forms of energy; its importance to the way we live cannot be over emphasized. It can be used to perform all kinds of useful functions and services that people desire, and at the point of use it produces no pollution. Electricity is a secondary source of energy, as it must be produced from primary sources such as coal, natural gas, oil, nuclear, hydro and, of late, increasingly from wind and solar. In 2010, the global production of energy was 21,325 trillion kWh, which is equivalent to 77 EJ. However, since 68% of electricity is derived from burning fossil fuels at an average efficiency of 38%, the total primary energy the world consumed as

electricity amounts to 135 EJ, or roughly 27% of the total energy consumed. Electricity demand in the world is projected to rise at a rate greater than for total energy, and meeting that need for additional sources for producing electricity is critical human welfare.

Several entries in this section are devoted to the production of electrical power. As the fuel with the lowest levelized cost of electricity, coal remains the dominant fuel for producing electricity, and furnishes over half of the electricity. It results in the emission of about 900 g CO₂ per kWh, highest of any fossil resource. The development of coal power in future is going to be constrained by the concerns of CO₂ emissions, and Beer ([► CO₂ Reduction and Coal-Based Electricity Generation](#)) discusses the different technology options for increasing the efficiency of coal-fired power plants and reducing the CO₂ emissions. Ultra supercritical steam cycles and combined heat and power applications are relatively straightforward to apply, and have the potential to reduce the largest tonnage of emissions.

Of course, there are other pollutants also that are emitted when coal is used for power productions. Chief among them are oxides of sulfur and nitrogen and toxic metals such as mercury, selenium, and arsenic, as well as carcinogenic PAH-containing soot aerosols. Moyeda ([► Pulverized Coal-Fired Boilers and Pollution Control](#)) reviews the technologies for scrubbing the stack gas emissions. The deployment of scrubbers for the nitrogen and sulfur oxides following the Clean Air and Water Act in the 1980s greatly improved the quality of air and water systems.

Natural gas is the other major fuel for producing electric power. When used in a gas-fired combined-cycle mode, it produces only 380 g CO₂/kWh. Smith and Gülen ([► Natural Gas Power](#)) review the technology for power generation from natural gas. They discuss the different sources of natural gas, provide a brief history followed by detailed thermodynamics of gas turbines, and modern power plants designs that incorporate combined cycle for maximum overall efficiency. They also review emissions from NGCC plants, notably NO_x, and strategies for minimizing them. The entry concludes with a discussion of power plant economics.

Carbon capture and sequestration (CCS) as a strategy for stabilizing atmospheric CO₂ levels is receiving considerable attention. Friedman ([► CO₂ Capture and](#)

Sequestration) provides an overview of the state of technology and the different embodiments of CCS. These cover various strategies for first capturing the CO₂, which may be performed pre- or post-combustion, or from the atmosphere. He discusses the energy requirements under these different scenarios. The entry also reviews the geochemistry of various options for sequestering the captured CO₂, either in deep saline aquifers or in other geological formations.

Bibliography

1. Meadows D, Randers J, Meadows D (2004) Limits to growth, the 30-year update. Chelsea Green Publishing, White River Junction, Vermont
2. BP statistical review of world energy, June 2011 From www.BP.com
3. Energy Information Administration, Annual energy review 2010. http://www.eia.gov/emeu/aer/pecss_diagram.html

Fuel Cell Comparison to Alternate Technologies

JULIA KUNZE^{1,2}, ODYSSEAS PASCHOS², ULRICH STIMMING^{1,2}

¹Institute for Advanced Study (IAS), Garching, Germany

²Department of Physics, Technische Universität München, Garching, Germany

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Classical Heat Engines

Electrochemical Systems

Electromobility: An Example for Combining Key Energy Technologies

Future Directions

Bibliography

Glossary

Fuel cell A fuel cell is an electrochemical cell that can convert the chemical energy stored in a given fuel into electrical energy.

Electrochemical capacitor An electrochemical capacitor (supercapacitor, ultracapacitor or double-layer capacitor) is an electrochemical device that can store and convert energy by charging/discharging the electrochemical double-layer of 2 electrodes with large surface areas and thus large double layer capacitances.

Battery A battery or Voltaic cell consists of one or more electrochemical cells which store and convert chemical energy into electric energy.

Ragone plot A Ragone Plot compares the performances of different energy storing devices by plotting power densities or specific power [W/kg] versus energy densities or specific energy [Wh/kg].

Electromobility Electromobility is a mobility concept in which electric vehicles instead of vehicles powered by internal combustion engines are used.

Definition of the Subject and Its Importance

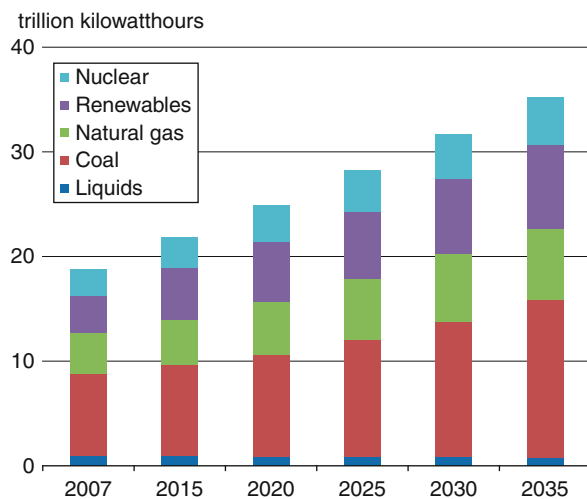
The actual energy demand and consumption issues make it necessary to critically discuss and compare different energy conversion and storage systems. At present, only one third of the primary energy is converted into end energy, for example, electrical energy. Losses are associated with a high consumption of fossil fuels and large CO₂ emissions. They can be avoided by considering important electrochemical processes for energy conversion, using batteries, fuel cells, supercapacitors and electrochemical photovoltaics and by incorporating energy storage, employing rechargeable batteries, supercapacitors, generation of hydrogen via electrolysis, and generation of methanol.

Fuel cells can offer clean power generation and have the potential to convert fuels directly into electrical energy with high efficiencies. Today, however, they cannot compete with heat engines because of much higher costs, inferior power performance, and insufficient durability and lifetime. As yet, no single electrochemical power source can match the characteristics of the internal combustion engine. A competitive system in comparison to heat engines can be envisioned when available electrochemical power systems are combined. In such hybrid electrochemical power systems, batteries and/or supercapacitors would provide high power and the fuel cells would deliver electricity with high efficiency. The consequence would be to considerably reduce costs of the electrochemical systems.

Moreover, when we compare different systems, we face the genuine problem of geometry and thus the costs related to the space needed for the processes involved. The latter explains the so far success of thermodynamic-based systems (e.g., heat engines) compared to alternative energy technologies (e.g., fuel cells). This is also evident when the power output of a coal power plant which lies in the GW range is compared to the power that a fuel cell can deliver which lies in the MW range. A possible strategy would be to develop highly efficient fuel cell systems that can push their power to larger outputs. Current research is aiming at increasing the power output of a fuel cell without compromising its efficiency; however, the advantages of conventional systems cannot be ignored.

Introduction

Our present situation concerning major (fossil) energy carriers and energy consumption is characterized by limited reserves and resources as well as emission problems, considering crude oil, natural gas, coal, and uranium as primary energy sources. At the same time, the worldwide demand for electrical energy has increased from 8.3 million GWh in 1980 to 18.9 million GWh in 2006 and is estimated to further increase up to 30.7 million GWh in 2030 (see [1] (Fig. 1)).



Fuel Cell Comparison to Alternate Technologies.

Figure 1

Worldwide demand for electrical energy (From [1])

The problem of the energy management is of a rather complex nature. Increasing contributions of renewable energies such as wind power, solar power, and wave power tend to complicate the grid management. Therefore, generation of electrical energy is only part of the challenge. The management and storage of electrical energy will become essential in order to maintain the present grid quality [2].

Energy conversion processes today are under special consideration because of two issues associated with them:

1. The limited availability of primary energy carriers
2. The emission of pollutants with local and global negative effects on the environment

On average, energy conversion processes aiming at generating electrical power reveal efficiencies not much higher than 30%. This indicates that losses in the form of heat or chemical substances amount to more than two thirds of the primary energy. Conventional processes, for example, those in a heat engine-based power plant, are volume processes such as combustion, which results in mechanical and then in electric energy. By contrast, other technologies such as photovoltaics, batteries, fuel cells, or supercapacitors are based on interfacial transfer of energy and/or charge. In contrast to the performance of heat engines which is limited by the Carnot efficiency, interfacial reactions are usually of much higher thermodynamic efficiency.

Classical Heat Engines

Thermodynamic considerations lead to the first heat engines that were used to generate mechanical or electrical energy. Since the eighteenth century, conventional reciprocating steam engines have served as mechanical power sources, with notable improvements being made by James Watt. The first commercial central electrical generating stations in New York and London, in 1882, used steam engines [3]. These first-generation heat engines are still serving as power plants today.

Further developments lead to the introduction of combined cycle power plants. Here, usually a combination of several cycles, operating at different temperatures yields considerably higher system efficiency. Heat engines are only able to use a portion of

the energy, usually 35–41%. The remaining heat is generally wasted.

In a combined cycle power plant (CCPP), or combined cycle gas turbine (CCGT) plant, a gas turbine generator produces electricity and the waste heat is used to make steam for generating additional electricity via a steam turbine; this last step enhances the efficiency for electricity generation to about 60%, because the temperature difference between the input and output heat levels is higher leading to an increase in the Carnot efficiency. Most modern power plants in Europe and in North America are of this type.

If the waste heat of a conventional thermal power station is utilized for district heating, it is called cogeneration. This heat gives an additional efficiency of about 40–50% which leads to an increase of the overall efficiency to ~90%. The fuels used in these engines comprise black and brown coal and nuclear power.

In contrast to the stationary systems described above, mobile engines are needed in many energy production applications. The best known mobile engines are the four-stroke (Otto), the diesel, and the Stirling engine.

Heat engines have limited efficiencies which are determined by the Carnot cycle. Practical issues reduce the efficiency of steam engines, due to limits of convective heat transfer and viscous flow (friction). There are also mechanical considerations, for example, limitations imposed by the materials such as nonideal properties of the working gas, thermal conductivity, tensile strength, creep, rupture strength, and melting point.

Electrochemical Systems

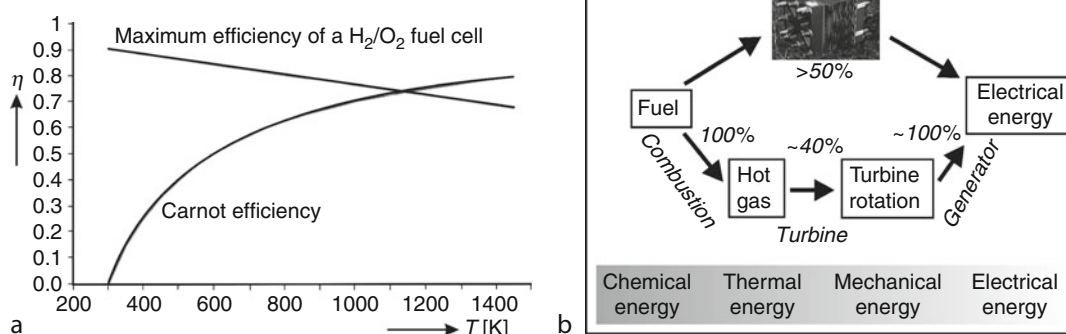
Energy Conversion and Storage

Alternatively to conventional Carnot-based heat engines, electrochemical systems have the potential to be employed to generate and store energy. Important electrochemical processes for energy conversion are considered in batteries, supercapacitors, electrochemical solar cells, and fuel cells, and energy storage is incorporated by employing rechargeable batteries, supercapacitors, generation of hydrogen *via* electrolysis, and generation of methanol from electrochemically generated hydrogen and CO₂/CO-containing synthetic gas. Electrochemical energy conversion systems are not subject to the Carnot cycle limitations and may operate with much higher efficiencies than combustion engines

and related devices. Heat engine cycle processes are volume processes, whereas all electrochemical systems that are used for energy conversion and storage are based on reactions that take place at an interface. This makes them dependent on the surface morphology and its physical and chemical properties.

A *fuel cell* is an electrochemical conversion device. It produces electricity from fuel on the anode side and an oxidant on the cathode side, which react in the presence of an electrolyte. Fuel cells can operate continuously as long as the necessary flows of reactants and reaction products are maintained. They are thus thermodynamically open systems. Many combinations of fuel and oxidant are possible. A hydrogen cell uses hydrogen as fuel and oxygen, usually from air, as oxidant. Other fuels include hydrocarbons and alcohols. Other oxidants include chlorine and chlorine dioxide. The principle of fuel cells is already known for 165 years. In 1838, Christian Friedrich Schönbein discovered that an electrical voltage is measured between two platinum wires in an electrolyte solution surrounded by hydrogen and oxygen. He published these results under the title “On the Voltaic Polarization of Certain Fluid and Solid Substances.” Sir William Robert Grove investigated the new effect intensely and was the first to develop useful fuel cells. He connected several elements in series and called these systems “gas batteries.” The further development of this concept to a really efficient electrical source proved to be so difficult that it took 100 years until it improved essentially. After the design of the first electrical dynamo by Werner von Siemens in 1867, electrical generators were energy sources that efficiently delivered electricity in almost unlimited amounts. This was the reason why fuel cells were practically replaced completely. *Fuel cells* have the potential for high conversion efficiencies of chemical energy into electrical energy; at the moment the attainable efficiency is up to 65%, depending on fuel and conditions. Fuel cells can substitute other technologies or be operated in synergy, for example, with combustion engines in cogeneration of electricity and of heat and cold, in residential or mobile applications.

Figure 2 shows a comparison of the thermodynamic efficiencies of a hydrogen/oxygen fuel cell and of a Carnot cycle. In the future, fuel cells might be able to convert the used fuels into electrical energy with efficiencies of >70%. The difference between the



Fuel Cell Comparison to Alternate Technologies. Figure 2

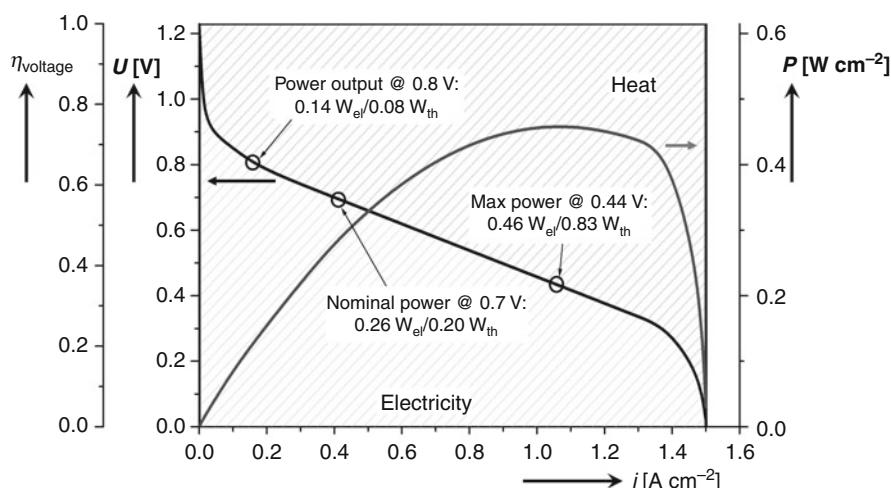
(a) Ideal thermodynamic efficiency of polymer electrolyte membrane fuel cells (PEMFCs) compared to that obtained in the Carnot process. (b) Comparison of processes in a cogenerated heat engine with fuel cell performance (From [2])

theoretical and practical energy storage capabilities is related to several factors, including (1) inert parts of the system such as conductive diluents, current collectors, containers, etc., that are necessary for its operation, (2) internal resistances within the electrodes and electrolyte and between other cell components, resulting in internal losses, and (3) limited utilization of the active masses, as, for example, parts of the fuel in a fuel cell leave the cell without reaction or as, for example, passivation of electrodes makes them partially electrochemically inactive. Fuel cells offer a very clean power generation. They are quiet in operation and can be located close to the application. They produce much less greenhouse emissions and can be more efficient in conversion of the energy in a fuel into power than gasoline engines or utility thermal power plants. Fuel cells are best utilized as a steady energy source and not as a power source to supply dynamic demands. For applications that require varying power demands, such as automotive propulsion, the use of the fuel cell in a hybrid configuration with a battery or an electrochemical capacitor is envisioned. The fuel cell provides steady power delivery while the battery or the electrochemical capacitor handles the surge for regenerative braking and acceleration as well as initial start-up. There is a *certain number of fuel cells under development*. Usually fuel cells are classified by the electrolyte that they are using. The electrolyte is a major component of the fuel cell since it determines the operating temperature, fuel, oxidant, poisons, catalysts, electrode design,

and system design. According to the type of electrolyte that is used, fuel cells can be classified into six major categories.

The polymer electrolyte membrane fuel cell (PEMFC) uses a solid-type electrolyte that typically cannot withstand high temperatures, and the relatively low operating temperature results in slow reaction rates. The operating temperature of a PEMFC is typically between 50°C and 100°C . The slow reaction rate is overcome by using highly active catalysts such as platinum. The fuel utilized is pure hydrogen and the electrical power that it produces is typically less than 250 kW. The fuel and oxidant must be free of carbon monoxide, which is a poison because it gets strongly adsorbed by the catalyst reducing the active surface area. The main applications for PEMFCs are for small stationary units and as power supplies in vehicles. The direct methanol fuel cell (DMFC) is similar to the PEMFC but methanol is used as a fuel instead of hydrogen. It can be used for portable applications such as laptops, cell phones, and MEMS devices. Maximum power of DMFCs is typically less than 10 kW.

Figure 3 shows a typical polarization curve of a low-temperature PEMFC. The curve is similar for all the other types of fuel cells as same trends are observed. At high operation voltages, the power output of a fuel cell is controlled mainly by the electrocatalytic properties of the catalyst. In this region (>0.8 V), the current density is controlled by the charge transfer and the dissipated heat is minimized. Challenges in fuel cell



Fuel Cell Comparison to Alternate Technologies. Figure 3

Conversion of chemical energy in fuel cells (voltage efficiency η_{voltage} versus current density i). Black line: cell voltage U ; gray line: power density P

catalyst research are to obtain high electric power density, high electric conversion efficiency, and low material costs. At high electric power density, mass transport is the dominant factor affecting the fuel cell operation. This region can be expanded by careful engineering design of the fuel cell system. A general research trend is to obtain as high current densities as possible at high operation voltages.

For *high electric conversion efficiency*, one should aim at the use of less cogenerated heat which would also allow for a simplification of the design and the need of less fuel (hydrogen) for the same electrical energy. To guarantee low material costs, less noble metal should be incorporated, because high catalyst utilization is needed, and highly efficient catalysts are needed for anode and cathode side. The goal in fuel cell research must therefore be a deeper understanding of the parameters controlling electrocatalytic activity which would enable us to propose rational structures for catalysts in fuel cells in the future.

Alkaline fuel cells (AFCs) were used to provide electrical power for many manned spacecrafts. The electrolyte is KOH and the catalysts include silver, nickel, and different metal oxides. AFC catalysts are relatively inexpensive compared to catalysts used for other types of fuel cells. The fuel and oxidant used in an AFC must be completely free of CO_2 since even a small

amount reacts strongly with the electrolyte, producing forms of carbonates that poison the ionic conductivity of the electrolyte. Therefore, pure hydrogen and oxygen must be used, limiting the use of the AFC to special applications, like spacecrafts and submarines, where cost of the fuel and oxidant is not a major issue.

The phosphoric acid fuel cell (PAFC) was the first fuel cell to be commercialized and shares some technologies with the PEMFC, such as the porous electrodes and the platinum catalysts. The liquid phosphoric acid allows high operating temperatures, around 200°C . Fuels must be free of carbon monoxide, as with the PEMFCs. With rated power over 50 kW, PAFC systems are used for stationary applications.

Molten carbonate fuel cells (MCFCs) operate at temperatures around 650°C . The electrolyte is a mixture of molten carbonate salts, such as lithium and potassium carbonate. At the operating temperature, the carbonates melt and become conductive to carbonate ions. Due to the high temperature, high reaction rates can be achieved with low-cost catalysts such as nickel. Rated power for MCFC systems is typically greater than 1 MW.

The last fuel cell system is the solid oxide fuel cell (SOFC). Typical operating temperature for the SOFC is between 600°C and $1,000^\circ\text{C}$. It has the same advantages of the MCFC (high reaction rates, inexpensive catalysts,

and natural gas can be used as a fuel). The electrolyte is in solid form, the most common kind being yttria (Y_2O_3) stabilized with zirconia (ZrO_2), which is a ceramic and therefore makes the fabrication process complex and difficult. The power output can be greater than 200 kW, making the SOFC ideal for large stationary applications, such as power stations.

High-temperature fuel cells such as *molten carbonate fuel cells* (MCFC) or *solid oxide fuel cells* (SOFC) deliver off heat at high temperature level ($T > 600^\circ\text{C}$). This can be used in high-temperature fuel cell cogeneration for district heating (heat extraction), for absorption refrigerators and heat pumps. As the energy demand in residential applications is complementary for heating and for cooling over the year, such a combination would be very useful.

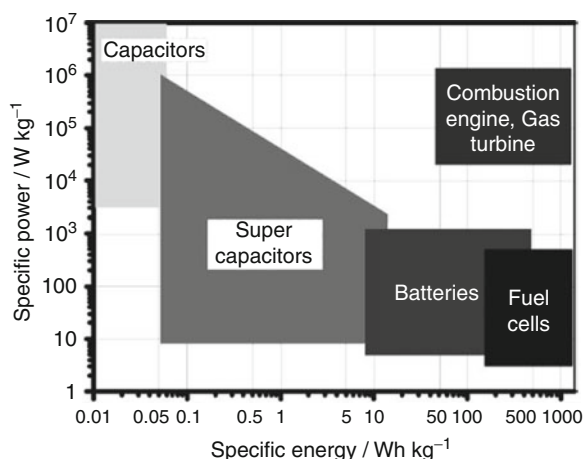
A *battery* or voltaic cell consists of one or more electrochemical galvanic cells which store and convert chemical energy into electric energy. Since the invention of the first Voltaic pile in 1800, the battery has become a common power source for many household and industrial applications. The name “battery” was coined by Benjamin Franklin for an arrangement of multiple Leyden jars, an early type of capacitor. There are different types of batteries which comprise non-rechargeable batteries (primary batteries) in which chemical energy is stored internally, the discharging reaction, that takes place at the electrode/electrolyte interface, is irreversible, and rechargeable batteries or accumulators (secondary batteries) having a reversible discharging reaction [4].

In the nineteenth century, no difference was made between *batteries* and *fuel cells*. Today the definitions have been refined: batteries convert and store electrical energy and hence represent a thermodynamically closed system; energy storage and conversion occur in the same compartment. Fuel cells are open systems where the anode and cathode are just charge-transfer media and the active masses undergoing the redox reactions are delivered from outside the cell, either from the environment, for example, oxygen from air, or from a tank, for example, fuels such as hydrogen and hydrocarbons. Energy storage (in the tank) and energy conversion (in the fuel cell) are thus locally separated.

Electrochemical capacitors (EC), also called supercapacitors, store energy using either ion adsorption or fast surface redox reactions. In the first case,

they are called *electrochemical double layer capacitors* (EDLC), in the second case they are called *pseudocapacitors*. They can complement or replace batteries in electrical energy storage and harvesting applications, when high power delivery or uptake is needed. Several types of ECs can be distinguished, depending on the charge storage mechanism as well as the active materials used. EDLCs, the most common devices at present, use carbon-based active materials with high surface areas. Pseudocapacitors or redox supercapacitors use fast and reversible surface or near-surface reactions for charge storage. Transition metal oxides as well as electrically conducting polymers are examples of pseudocapacitive active materials. Electrochemical capacitors currently fill the gap between batteries and conventional solid state and electrolytic capacitors. They store 100–1,000 times more charge than the latter due to a much larger surface area ($1,000\text{--}2,000\text{ m}^2\text{g}^{-1}$) available for charge storage in EDLC. They have a lower energy density than batteries, and this limits the optimal discharge time to less than a minute, whereas many applications clearly need more [5]. Since the early days of EC development in the late 1950s, there has not been a good strategy for increasing the energy density; only incremental performance improvements were achieved from the 1960s to 1990s. The increase in performance that has been demonstrated in the past few years is due to the discovery of new electrode materials and improved understanding of ion behavior in small pores, as well as the design of new hybrid systems combining faradic and capacitive electrodes [6].

To compare the power and energy capabilities of fuel cells, batteries, and supercapacitors, a representation known as the *Ragone plot* has been developed. The terms *specific energy* [Wh/kg] and *energy density* [Wh/L] are used to compare the energy contents of a system, whereas the rate capability is expressed as *specific power* [W/kg] and *power density* [W/L]. A simplified Ragone plot (Fig. 4) discloses that fuel cells can be considered to be high-energy systems, whereas supercapacitors are considered to be high-power systems. Batteries have intermediate power and energy characteristics. There is some overlap in energy and power of supercapacitors or fuel cells with batteries. Batteries with thin-film electrodes exhibit power characteristics similar to those of supercapacitors.



Fuel Cell Comparison to Alternate Technologies.

Figure 4

Simplified Ragone plot of the energy storage domains for the various electrochemical energy conversion systems compared to an internal combustion engine and turbines and conventional capacitors (From [7])

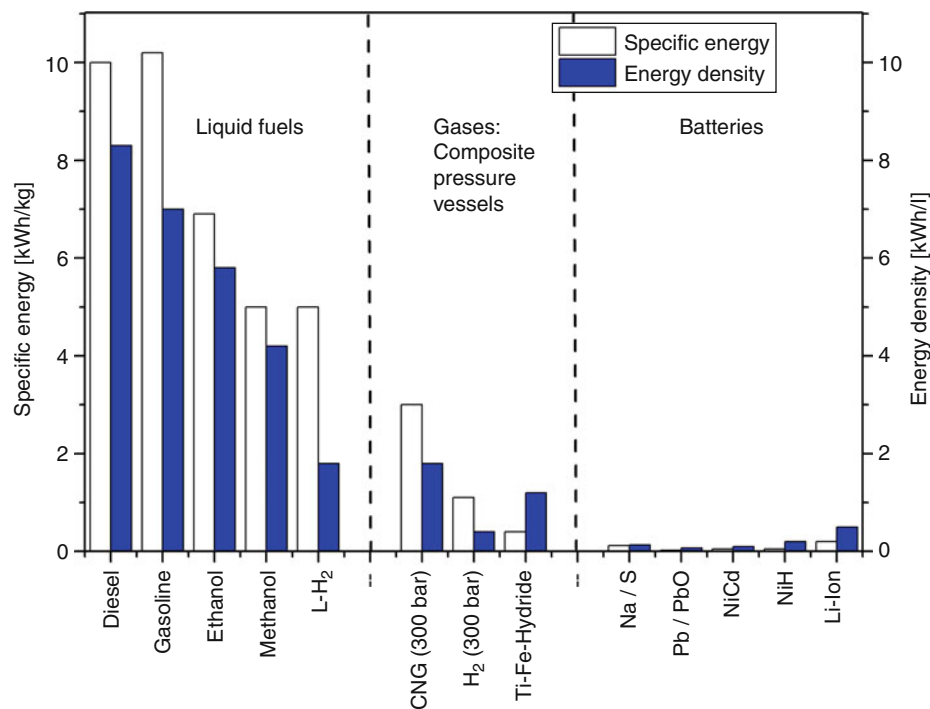
Not all systems can be described using these definitions, since there are also *combined systems* such as metal–air batteries [8–10] which contain a battery electrode (metal anode) and a fuel cell electrode (air cathode), or redox flow batteries [11, 12] which are a form of rechargeable battery in which electrolyte containing one or more dissolved electroactive species flows through an electrochemical cell. Additional electrolyte is stored externally, generally in tanks, and is usually pumped through the cell of the reactor. The best known example is the chromium–iron battery [13]. The active masses are represented by Fe^{3+} and Cr^{2+} ions. In these redox storage devices, the mechanical aging of the accumulator electrodes is prevented. Flow batteries can be rapidly recharged by replacing the electrolyte while simultaneously recovering the spent material. Due to the overlapping properties of batteries and fuel cells, it is nowadays considered to find a definition that recombines the two systems again, to no longer differentiate between batteries and fuel cells like in former times.

In order to bridge the gap between batteries and electrochemical capacitors, a capacitive or pseudocapacitive electrode can be combined with a battery electrode. Batteries exhibit usually high energy densities,

while capacitors can deliver high power at the cost of low energy storage. Research efforts have been focused to increase the gravimetric energy of electrochemical capacitor electrodes by using materials with enhanced gravimetric capacitances (gravimetric charge storage per volt) [14]. The latter can be achieved by using carbon subnanometer pores for ion adsorption or by incorporating the pseudocapacitance of nanostructured transition metal oxides. A recent report by S.W. Lee et al. [15] demonstrated that by using a novel class of electrodes for lithium storage high gravimetric energy ($200 \text{ Wh kg}_{\text{electrode}}^{-1}$) was delivered at a high power of $100 \text{ kW kg}_{\text{electrode}}^{-1}$. Those electrodes were based on functionalized multiwalled carbon nanotubes (MWNTs) that included stable pseudocapacitive functional groups and were assembled using a layer-by-layer technique. Therefore, it was shown that it is possible to benefit from both the capacitor and the battery properties.

A critical parameter of fuel selection is its energy density. The latter should be as high as possible. Additionally, the choice of fuel is also affected by the ease of transportation and storage which in general should be as easy and uncomplicated as possible. The choice of the fuels determines which problems, including environmental problems, will arise.

Figure 5 shows the theoretical specific energies [(kWh)/kg] and energy densities [(kWh)/l] of various rechargeable battery systems in comparison to fuels, such as gasoline, natural gas, and hydrogen. The inferiority of batteries is evident. These values are an indication for the maximum energy content of certain systems. The practical values are significantly lower than the theoretical ones. A rechargeable battery has a practical energy content which is 25% of its theoretical value while a primary battery normally yields more than 50% of its theoretical value. The theoretical efficiencies of fuel cells for the conversion of fuels into electrical energy are higher than 70%. A number of factors influence the observed differences between theoretical and practical energy storage capabilities. The systems contain inert parts such as conductive diluents, current collectors and containers, and internal resistances which results in internal losses and limited utilization of the active masses. Parts of the fuel can, for example, leave the fuel cell without having reacted or partial passivation of the electrodes can make parts of



Fuel Cell Comparison to Alternate Technologies. Figure 5
Theoretical specific energies [k Wh/kg] and energy densities [k Wh/l] of various rechargeable battery systems compared to fuels, such as gasoline, natural gas, hydrogen, and alcohols

the surface area inactive. For practical use, however, as batteries and fuel cells are not subject to the Carnot cycle limitations, they may operate with much higher efficiencies than combustion engines and related devices. The efficiencies of PEMFCs, batteries, and supercapacitors are summarized for comparison in Table 1.

Electrochemical photovoltaics are a specific type of photovoltaic systems to convert solar energy, that is, solar radiation into electrical energy. Today, there is a variety of solar cells already on the market. Difference must be made between electrochemical photovoltaic cells and conventional silicon (Si)–based photovoltaic cells; in electrochemical photovoltaic cells two interfaces exist at which charge transport switches from electronic to ionic and back. In photoelectrochemistry a photoactive semiconducting working electrode is used together with a metallic counter electrode. Both are immersed in an electrolyte containing a suitable redox couple. Upon radiation of the semiconductor–electrolyte interface with light having a higher energy

Fuel Cell Comparison to Alternate Technologies.

Table 1 Practical efficiencies of fuel cells, batteries, and supercapacitors

System	Process	Efficiency
PEMFCs	Electrolysis-H ₂ storage – FC operation	40%
Batteries	Charging – Discharging	80%
Supercapacitors	Charging – Discharging	95%

than the bandgap of the semiconductor, electron–hole pairs are formed and separated. Photogenerated majority carriers accumulate at the back side of the semiconductor, minority carriers travel to the semiconductor–electrolyte interface. The majority carriers are collected with a charge-collecting substrate and transported to the counter electrode where they

electrochemically react with the redox couple in the electrolyte. In the 1970s, wet-type electrochemical photovoltaic cells were discovered [16–18]. Michael Grätzel demonstrated that dye molecules can be adsorbed onto nanocrystalline TiO_2 working electrodes and thus extended the concept to dye-sensitized solar cells (DSSCs) [19–22].

It is noteworthy that electrochemical photovoltaic cells without dyes have working (semiconductor) and counter electrodes that are both immersed in the redox electrolyte. After the formation of electron/hole pairs, specific reactions occur at the semiconductor electrode and at the metal counter electrode. Charge balance is maintained via oxidation and reduction processes, and electrodes are often instable in aqueous media. Photoelectrochemical photovoltaic cells could only replace Si-based photovoltaics if a stable semiconductor material with a band gap of ~ 1.4 eV is found [23–26]. In dye-sensitized solar cells, a dye-sensitized porous nanocrystalline TiO_2 photoanode with a 1,000-fold surface area compared to flat electrodes has been used. Furthermore, the charge transport of the photogenerated electrons passing through all the particles and grain boundaries is highly efficient [27]. The conversion efficiency of DSSCs has improved to 11.5% [28] since it was first reported with an efficiency of 7.1% [26]. Their efficiencies are now comparable with amorphous silicon solar cells [29] but at a much lower cost.

Electromobility: An Example for Combining Key Energy Technologies

Concerns about global warming and the depletion of petroleum resources demand a shift in automotive transportation from petroleum-based internal combustion engines to electrically driven vehicles (EVs). In the concept of electromobility, all the above-mentioned technologies should be united and combined in order to synergistically provide power to an EV.

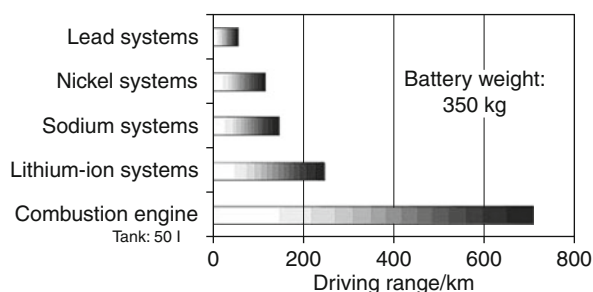
Hybrid electric vehicles are a first step, where an internal combustion engine is assisted by a battery-driven electric motor, enabling increased fuel efficiencies when using high-power and low-energy (ca. 1 kWh) batteries. Further reduction or even elimination of petroleum-based fuels can be achieved by successively increasing the onboard stored electrical energy, thus

enabling plug-in hybrid electric vehicles, short-range electric vehicles and, ultimately, long-range electric vehicles. Such a development would be the perfect solution to realize electric traction in general.

Together with batteries and supercapacitors, fuel cells are a key component that must be investigated in order to build electric vehicles suitable for continuous use, especially in public transport. Fuel cells have the potential to offer clean power generation, are quiet in operation, and can be located close to the application. They produce much less greenhouse emissions and can be more efficient in conversion of the energy in a fuel into power than gasoline engines or utility thermal power plants. Fuel cells are best utilized as a steady energy source and not as a power source to supply dynamic demands. For applications that require varying power demands, such as automotive propulsion, the use of the fuel cell in a hybrid configuration with a battery or an electrochemical capacitor is required. The fuel cell provides steady power demand while the battery or the electrochemical capacitor handles the surge for regenerative braking and acceleration as well as initial start-up. This makes fuel cells an essential component for hybrid vehicles with adequate ranges as they are required for public transport and goods traffic.

Fuel cells, batteries, and supercapacitors have a similar configuration, that is, two electrodes separated by an electrolyte. The latter are designed for high power and long life service. Currently nonetheless, the status of batteries, supercaps, and fuel cells does not allow for a satisfactory use in electric vehicles due to tremendous weight and safety issues, too high costs, and too short ranges. Thus, fundamental research is necessary to develop novel battery, supercapacitors, and fuel cell concepts for a new generation of electric vehicles.

Nowadays, also concepts are developed in which photovoltaic cells integrated on the roof or in the windows of EVs can be used to recharge the batteries. It is known from literature that EVs suffer from too short ranges and too high weight [7, 30]. Figure 6 shows typical driving ranges of battery-powered cars compared to ranges of cars powered by modern combustion engines. As it can be seen, for EVs powered with any secondary battery system (i.e., lead acid, nickel metal hydride, sodium sulfur, and lithium ion), the driving ranges that they possess are inferior



Fuel Cell Comparison to Alternate Technologies.

Figure 6

Comparison of the driving ranges for a vehicle powered by various battery systems or a gasoline-powered combustion engine (From [7])

compared to those of conventional combustion engines. Therefore, it is illustrated clearly why rather fuel cells and not batteries should be considered for replacement of combustion engines.

Fuel cells and batteries can be used to build full electric vehicles and thus provide pure electrification solutions for different mass and usage segments of automotive application. Battery powered EVs based on lithium ion technology will be limited to small-vehicle low-mileage-per-day applications which is due to the low specific energy and long charging times. The required 300 mile range has been reached with hydrogen fuel cell vehicles and the latter have also been proven to be able to operate in all climates. Nevertheless, the costs of the Pt-based catalysts needed in a fuel cell are too high, the hydrogen infrastructure is not present and cost-effective renewable sources of primary energy are not yet efficiently used. Drastic decreases in the amount of Pt used are required which is put into practice via a number of strategies currently under development [30].

A very promising approach to energy conversion in electric vehicles is the use of metal–air batteries, for example, Li–air systems. As was discussed in the earlier section, the metal–air battery combines fuel cells with batteries. In the latter example, substitution of the conventional lithium intercalation cathode with an air cathode is projected to increase the cathode specific capacity by a factor of ~ 3.5 . Additionally, the concept of Li–air systems has another major advantage; it can

drastically decrease the weight of the battery pack. Specifically, for the case of short-range EVs (~ 25 kWh), the total battery would weigh only ~ 40 kg. For the case of a full-range EV, the projected Li–air battery-pack weight is only 120 kg. This is significantly smaller compared to the minimum weight of 330 kg for a best case Li-ion battery pack. Therefore, the Li–air battery technology would be a very promising and enabling technology for full-range EVs. Li–air batteries give promise of extending the range, but scientists and engineers will have to face many challenges in order to prove the research in this area effective.

The electrical energy required on board to travel a certain range before refueling is provided by the energy storage/generation system and is determined by the total gross energy consumed minus the energy recovered from regenerative braking ($Wh_{net}/mile$). All loads on board can be major drains of energy since they are powered by the electrical storage/generation system which affects the driving range. Loads that severely affect the driving ranges are heating and cooling devices in the car. The vehicle mass (m) including the mass of the powertrain system and of the passengers, the frontal area (A), the drag coefficient (C_d), and the drive cycle are elements that influence the $Wh_{net}/mile$ requirement to the electric motor the most. Furthermore, the rolling resistance and the ancillary loads mentioned above have strong influence.

An equation has been developed to correlate the different parameters [30]:

$$\text{Energy required} = a_1 \times m + a_2 \times C_d \times A + a_3$$

a_1 , a_2 , and a_3 are obtained from modeling results. The model is primarily based on the greenhouse gases, regulated emissions, and energy use in transportation (GREET) software developed by the Argonne National Lab [31]. In Table 2 [30], estimates are listed that start from a small subcompact car ($C_d \times A = 0.6 \text{ m}^2$) with a total vehicle mass of 1,200 kg for which the energy required to drive the Environmental Protection Agency (EPA) city or highway cycle is 138 or 156 $Wh_{net}/mile$, respectively. The energy required for a van ($C_d \times A = 1.56 \text{ m}^2$) weighing 2,500 kg is 283 or 361 $Wh/mile$ for city or highway cycle, respectively. The energy requirement for driving in the city is lower because the vehicle has less resistance to overcome at lower

Fuel Cell Comparison to Alternate Technologies. Table 2 Electrical energy required as a function of the vehicle type and mass (analysis assumes 70% recovery of braking energy through regenerative braking)

Vehicle type	Mass (kg) ^a	$C_d \times A$ (m ²)	City (Wh _{net} /mile)	Highway (Wh _{net} /mile)
Sub-compact	1,200 (2)	0.60	138	156
Compact car	1,400 (4)	0.64	154	171
Midsize car	1,550 (4)	0.67	166	183
Full-size car	1,700 (5)	0.71	178	194
Minivan	2,200 (7)	0.93	224	250
Van	2,500 (8)	1.56	283	361

Taken from Ref. [30]

^aThe mass includes 65 kg passengers weight shown in parentheses

speed, electricity generation is more efficient at lower loads, and the use of regenerative braking preserves these advantages for the city cycle [30].

The above given examples show that the range problem is the most severe problem of batteries used to power electric vehicles. Fuel cells have the potential to overcome this issue. On one hand, rather cheap materials are used in Li ion batteries, while on the other hand they require a very large surface area of very finely controlled thin layers, interfaces, and separators. Li ion batteries are already mass produced for use in portable electronic devices and thus, many of the opportunities for cost reduction through scale have already been taken. Nowadays, the development and implementation of improved materials must be the way to further reduce costs.

In fuel cells, however, more expensive materials are employed even though their amount can be tremendously reduced. The total geometric surface area of the cells shows a 30-fold decrease due to higher current densities obtained with more conductive fuel cell electrolytes. Fuel cells can be built in a bipolar construction with cells stacked in series with the negative current collector of one cell serving as the positive current collector of the adjacent cell. Fuel cells require a hydrogen tank and an air compressor which makes their balance of plant more complex.

Strong investment is necessary to develop EVs based on the concept of electromobility. The advantages of all available energy conversion and storage technologies need to be utilized in order to achieve

efficient full vehicle electrification. It is very likely that both battery and fuel cell technologies will be developed for automotive application, since batteries are well suited for the use in small short-range vehicles and fuel cells are able to deliver electric power for long-range vehicles. Super capacitors can aid with their high power density output whenever it is needed. Photovoltaics integrated in the EV can convert the solar energy and charge the batteries. The main point is that until now no single technology shows the potential to monopolize the implementation into an electrical vehicle. For fuel cells specifically, research needs to focus into identifying novel catalyst materials that can reduce the amount of expensive catalysts currently used (e.g., Pt). Additionally, fuel cell technologies based on alternative fuels other than hydrogen need to be developed. For all suited energy technologies applicable in EVs, the key challenge is their smart combination in order to build long-range vehicles which produce less greenhouse gas emissions than cars powered by combustion engines.

Future Directions

In summary, even though the world is becoming aware of an upcoming energy crisis, according to all statistical data currently available, most of the worldwide energy demand is still delivered via fossil fuels, while renewables unfortunately still account only for a small part. It is a fact that use of fossil fuels results in greenhouse gas emissions to the atmosphere. For this reason, the

industrialized states, under the Kyoto protocol, have committed themselves to reduce their greenhouse gas emissions by a total of 5% in the period of 2008–2012 compared to 1990.

Most of the CO₂ emitted is a result of the inefficient conversion and distribution systems that we use. The current system efficiencies for energy conversion and distribution allow approximately only 30% of the primary energy to be delivered as usable energy. If we manage to increase the system efficiency, CO₂ emissions will be reduced drastically. Therefore, there is the urgent need to develop efficient energy conversion and distribution systems that emit either no or very little CO₂. A technology that has the potential to help in the reduction of CO₂ emissions is fuel cells.

Fuel cells have the potential to efficiently convert chemical energy to electricity. Even though as discussed the concept of fuel cells is not new and research efforts worldwide have been devoted to advance the technology, there are still many ways that need to be covered before bringing them into the market. Some of the key issues related to fuel cells that need to be tackled before even discussing about a broad commercialization of the technology are:

- (a) Low efficiency at high power densities
- (b) Use of hydrogen which is only a secondary energy carrier
- (c) Large amount of noble metal required for widespread applications
- (d) Cost and availability of Pt which is commonly used as catalyst
- (e) Durability and lifetime of the system

The first issue is the low efficiencies at high power densities of current fuel cell systems. For example in PEMFCs, typical power densities delivered are $\sim 1 \text{ W cm}^{-2}$ at 0.7 V operating voltage. That accounts for a voltage efficiency of $\sim 56\%$ while the overall system efficiency is even lower. The power output at the high voltage regime is controlled by electrocatalytic properties and suffers from losses attributed mainly to the slow kinetics of the oxygen reduction reaction. In order to achieve a high-power, high-efficiency system, it is necessary to identify and develop highly active catalyst materials that could push the power output to $\sim 1 \text{ W cm}^{-2}$ at 0.85–0.9 V.

The second issue relates to hydrogen which is the employed fuel for low-temperature fuel cells. The advantage of using hydrogen is that water is the only product. Nonetheless, we need to take into consideration that hydrogen is only a secondary energy carrier which is not freely available and is currently produced by “non-clean” methods, for example, reforming of natural gas. Hydrogen can be produced by regenerative sources but those suffer from low efficiencies. It would be beneficial to fabricate systems that can directly convert alternative fuels such as bio-derived alcohols which have higher energy densities and can be produced efficiently. Direct methanol fuel cells are a starting point but at their current status not a satisfying solution. Bio-ethanol is a very attractive candidate. It can be produced by cellulosic biomass, has high energy density, is nontoxic, and its liquid form allows using the current fuel infrastructure. Research should also aim at identifying highly active catalysts toward ethanol oxidation and electrolytes with low ethanol permeability that will result in designing an efficient system that will directly convert ethanol to electricity.

The last three issues refer to the current material systems used in fuel cells. Pt supported on carbon (Pt/C) is commonly used in PEMFCs as catalyst and in order to achieve desirable power densities high amounts of noble metal are required. Even though Pt possesses high catalytic activity, it is also associated with limited resources, and high price. Efforts to reduce the metal loading and alloy Pt with other inexpensive and more abundant elements, that is, metals, rare earths have shown improvements, however, there is no system found that would lead PEMFCs to a broader commercialization. Carbon is commonly used as support material in several fuel cell systems since it is inexpensive with high electronic conductivity and high availability. Carbon though is the main reason of the low lifetime since it undergoes corrosion. Efforts to nanostructure carbon, that is, nanotubes, nanofibers, have shown only unsubstantial improvements. Pt and C are materials that have been extensively investigated for many years but have not brought fuel cells to the market. Therefore research should focus to identify novel catalyst/support systems that would replace Pt and C.

Two possible future scenarios for stationary electricity production and road traffic can be envisioned.

For the first scenario small units (<1 MW) based on SOFC with natural gas, biogas, or ethanol could be implemented. In niche markets, electricity storage could be achieved with electrolysis to produce hydrogen that could subsequently be stored and then used in a fuel cell system. For larger units (>10 MW), SOFCs could be used with natural gas, biogas, carbon from biomass. For the latter, a negative CO₂ effect is possible (electricity will decrease CO₂ concentration in the atmosphere). For the second scenario of road traffic, two options are possible. City travel could be handled with battery-operated motor bikes and small cars, possibly in combination with a small (1–5 kW) fuel cell. Long-range travel could be achieved with larger vehicles using mid-temperature (200–300°C) bioethanol fuel cells hybridized with batteries. In both systems, supercapacitors could be employed to handle the high power needs. For both scenarios, all other energy conversion technologies (e.g., solar, wind) can be used in order to convert and subsequently store energy (i.e., battery recharging).

In conclusion, higher targets need to be set for the future and achieved before fuel cells can be a commercially viable technology. There have been numerous efforts to improve fuel cells but all solutions proposed so far are only small steps that address partially only one of the above-mentioned issues. High-risk, high-profit solutions are required that will drastically improve the current fuel cell status and will lead them to their broad commercialization. In fuel cell development, we need to match the available energy source with the application. Hydrogen does not seem to be an option. A direct conversion of an available fuel in a fuel cell would be more desirable. For example, alcohol oxidation at all temperatures and (hydro-) carbons could be used at high temperatures. In all situations, fuel cell research has to acknowledge the fuel problem. Therefore, if one believes that we are close to the end of research for fuel cells and other energy technologies then the reality is different ... we are only at the beginning.

Bibliography

- DOE/EIA-0484 (2010) National Energy Information Center, EI-30, U.S. Energy Information Administration, Forrestal Building, Washington, DC 20585
- Kunze J, Stimming U (2009) Electrochemical versus heat engine energy technology: a tribute to Wilhelm Ostwald's visionary statements. *Angew Chem Int Ed* 48:9230–9237
- Black WZ, Hartley JG (1985) *Thermodynamics*. Harper & Row, New York, pp 339–429
- Armand M, Tarascon J-M (2008) Building better batteries. *Nature* 45:652–657
- Kötz R, Carlen M (2000) Principles and applications of electrochemical capacitors. *Electrochim Acta* 45:2483–2498
- Simon P, Gogotsi Y (2008) Materials for electrochemical capacitors. *Nat Mater* 7:845–854
- Winter M, Brodd RJ (2004) What are batteries, fuel cells, and supercapacitors? *Chem Rev* 104:4245–4269
- Jacoby M (2010) Rechargeable metal-air batteries. *Chem Eng News* 88:29–31
- Cairns EJ, Albertus P (2010) Batteries for electric and hybrid-electric vehicles. *Annu Rev Chem Biomol Eng* 1:299–320
- Girishkumar G, McCloskey B, Luntz AC, Swanson S, Wilcke W (2010) Lithium-air battery: promise and challenges. *J Phys Chem Lett* 1:2193–2203
- Zhang HM, Zhang Y, Liu ZH, Wang XL (2009) Redox flow battery technology. *Prog Chem* 21:2333–2340
- De Leon CP, Frias-Ferrer A, Gonzalez-Garcia J, Szanto DA, Walsh FC (2006) Redox flow cells for energy conversion. *J Power Sources* 160:716–732
- Hagedorn NH (1983) The iron-chromium redox battery. *Abstr Pap Am Chem Soc* 186:28
- Chmiola J, Yushin G, Gogotsi Y, Portet C, Simon P, Taberna PL (2006) Anomalous increase in carbon capacitance at pore sizes less than 1 nanometer. *Science* 313:1760–1763
- Lee SW, Yabuuchi N, Gallant BM, Chen S, Kim B-S, Hammond PT, Shao-Horn Y (2010) High-power lithium batteries from functionalized carbon-nanotube electrodes. *Nat Nanotechnol* 5:531–537
- Gerischer H, Tributsch H (1968) Electrochemische Untersuchungen zur spectraleu sensibilisierung von ZnO-Einkristallen. *Ber unsenges Phys Chem* 72:437–445
- Hauffe K, Danzmann HJ, Pusch H, Range J, Volz H (1970) New experiments on the sensitization of zinc oxide by means of the electrochemical cell technique. *J Electrochem Soc* 117:993–999
- Myamlin VA, Pleskov YV (1967) *Electrochemistry of semiconductors*. Plenum, New York
- Gratzel M (2003) Applied physics: solar cells to dye for. *Nature* 421:586–587
- Gratzel M (2004) Conversion of sunlight to electric power by nanocrystalline dye-sensitized solar cells. *J Photochem Photobiol A* 164:3–14
- Hagfeldt A, Gratzel M (1995) Light-induced redox reactions in nanocrystalline systems. *Chem Rev* 95:49–68
- Hagfeldt A, Gratzel M (2000) Molecular photovoltaics. *Acc Chem Res* 33:269–277
- Ellis AB, Kaiser SW, Wrigton MS (1976) Visible light to electrical energy conversion. Stable cadmium sulfide and cadmium selenide photoelectrodes in aqueous electrolytes. *J Am Chem Soc* 98:1635–1637

24. Ellis AB, Bolts JM, Wrighton MS (1977) Characterization of n-type semiconducting indium phosphide photoelectrodes. *J Electrochem Soc* 124:1603–1607
25. Hodes G, Manassen J, Cahen D (1976) Photoelectrochemical energy conversion and storage using polycrystalline chalcogenide electrodes. *Nature* 261:403–404
26. Miller B, Heller A (1976) Semiconductor liquid junction solar cells based on anodic sulfide films. *Nature* 262:680–681
27. Wurfel U, Peters M, Hinsch A (2008) Detailed experimental and theoretical investigation of the electron transport in a dye solar cell by means of a three-electrode configuration. *J Phys Chem C* 112:1711–1720
28. Chen C, Wang M, Li J, Pootrakulchote N, Alibabaei L, Ngoc-le C, Decoppet J-D, Tsai J-H, Grätzel C, Wu C-G, Zakeeruddin SM, Grätzel M (2009) Highly efficient light-harvesting ruthenium sensitizer for thin-film dye-sensitized solar cells. *ACS Nano* 3:3103
29. Gratzel M (2007) Photovoltaic and photoelectrochemical conversion of solar energy. *Philos Trans R Soc A* 365:993–1005
30. Wagner FT, Lakshmanan B, Mathias MF (2010) Electrochemistry and the future of the automobile. *J Phys Chem Lett* 1:2204–2219
31. Wang MQ (2007) Greenhouse gases, regulated emissions, and energy use in transportation (GREET). The Argonne National Laboratory, Argonne. http://www.transportation.anl.gov/modeling_simulation/GREET/index.html

Fuel Cell Powered HEV Design and Control

ETIM U. UBONG¹, JIM GOVER²

¹Center for Fuel Cell Research and Powertrain Integration, Electrical & Computer Eng, Kettering University, Flint, MI, USA

²Department of Electrical Engineering, Kettering University, Flint, MI, USA

Article Outline

Glossary

Definition of the Subject

Introduction of Fuel Cells

Candidate Fuel Cells for Hybridization

Hybrid Technology Using Gasoline or Diesel, Fuel Cells, Ultra-Capacitors, and Batteries

Ultra-Capacitors

Conclusion

Bibliography

Glossary

Balance of plant (BOP) Auxiliary components that complement a power unit as a functional or operational unit.

Electric vehicle (EV) A vehicle that uses battery or any other onboard electrical power source to operate the vehicle.

Fuel cell A device that uses fuel and oxidant in an electrochemical reaction to generate direct electricity.

High-temperature proton exchange membrane fuel cell (HT-PEMFC) A PEM fuel cell that operates between 120°C and 200°C.

Hybrid electric vehicle (HEV) A vehicle that uses an auxiliary electrical power source, such as a fuel cell or a battery, along with an internal combustion engine (ICE) for propulsion.

Internal combustion engine (ICE) A device that converts the chemical energy of fuel in a cylinder to mechanical work through the piston, connecting rod and crank shaft mechanisms.

Low-temperature PEM fuel cell (LT-PEMFC) A PEM fuel cell that operates between 50°C and 80°C.

Proton exchange membrane (PEM) A fuel cell that uses either a sulphonated (low temperature) membrane or polybenzimidazole (high temperature) membrane to generate direct electricity.

Reformate A fuel derived from fossil fuel (natural gas) after reformation that is a mixture of hydrogen, carbon monoxide, and carbon dioxide.

Solid polymer fuel cell (SPFC) Same as PEM above.

Definition of the Subject

A hybrid vehicle implies a vehicle that is powered by more than one source of power. It can be an internal combustion engine and a battery, an ICE and fuel cell, and ICE and solar cells or fuel cell and batteries. The current trend of hybridization of internal combustion engines (ICE) is a result of the inefficiency of ICE. Spark ignition internal combustion engine, SI-ICE, is about 25–27% efficient, while compression ignition internal combustion engine, CI-ICE, (diesel) is 34–38% efficient [1, 2]. The remaining percentage of the input energy into ICE is expended in the form of heat loss to the coolant and exhaust, by convection, conduction, and radiation, and by combustion

inefficiency. The electric drive has a higher efficiency around 90%, as such, it is justifiable to incorporate electric drive into the hybridized powertrain. With the increased use of electronics in cars for various applications, such as visual display of GPS, fuel economy, maintenance schedule and diagnostics, voice activation of devices and DVD players, etc., there is a need to use additional power sources to maintain this level of power demand without compromising the power required for ignition, and operate the electronic control unit (ECU), anti-brake systems (ABS), and the fuel injection units.

In the hybridized ICE-powered vehicle, in the series mode of connection, the mechanical drive shaft that transmits power from the crankshaft through the gearbox to the wheels is replaced with a generator and electric wires to the electric motors spinning the wheels. The design configuration of the ICE-battery hybrid presents the base design of other complex hybrid vehicles architecture. The design structure of the fuel cell-powered hybrid electric vehicle includes fuel cell as the primary power source and a battery as a secondary power source for start-up, lighting, and power electronics support. The power transmission shaft reminiscent of the internal combustion engines (ICE)-powered vehicle is absent, rather electric motors spin the four wheels. Fuel cells generate direct electricity that is used to drive the electric motors in hybrid vehicles. As such, reliability is enhanced as there are fewer rotating components. The complexity and maintenance are reduced in electric machinery compared with mechanically driven machinery. The following report will present different types of fuel cells and their electrolytes, sources of fuels for fuel cells, those applicable to the transportation sector and the hybrid vehicle architecture. The ICE hybrid will also be presented as a base model, since its operation is well understood by the general public.

Introduction of Fuel Cells

Fuel cells are electrochemical electricity-generating devices that oxidize the fuel at the anode to electrons and protons and reduce oxygen at the cathode to produce heat and water. The splitting of hydrogen into protons and electrons is done at the catalyst layer (anode).

The electrons move across the external circuit generating electricity, while the protons permeate through the PEM to the cathode. The proton exchange membrane fuel cells are of interest in the mobility sector as it is modular (scalable) and, as such, can be built to desirable sizes without compromising the efficiency or performance. Various petroleum and nonpetroleum fuels, chemical compounds, and hydrogen carriers can be used as fuels in fuel cells after preprocessing. However, hydrogen is the most preferred fuel with the end products, in case of the proton exchange membrane (PEM) membrane being water, with the release of heat. Hydrogen is widely used in most fuel cells ranging from the low to the high-temperature PEM to solid oxide fuel cells (SOFCs).

Commonly Used Fuels in Fuel Cells

Fuels commonly used in fuel cells are: hydrogen, methanol, potassium hydroxide; phosphoric acid; methane (natural gas); carbon monoxide; carbonate salts of sodium, lithium, and potassium; and yttrium-stabilized zinc oxide as the electrolyte. [Table 1](#) summarizes various fuels and applicable fuel cells that utilize them.

Low-Temperature and High-Temperature PEMs

The low temperature PEM fuel cell shown in [Fig. 2a](#) relies on membrane humidification for ion transport from the anode to the cathode through Grotthuss mechanism or ion-hopping mechanism. If the membrane is not well humidified or is dry, vehicular mechanism is involved in the ion transport. For the high-temperature PEM fuel cell based on polybenzimidazole (PBI) membrane, which is presented in [Fig. 2b](#), the membrane is treated in sulfuric or phosphoric acid. Protons are transported in an acidic medium when the temperature is raised up to 130°C and above. The membrane performs much better at an elevated temperature of 170–200°C. Precautionary measures require that the membrane be raised to above 130°C before the load is applied. Better results are obtained if the cell is allowed to run and equilibrate at a reduced load for about 100 h before readings are taken. This allows the membrane to equilibrate. During the shutdown period, the membrane should be purged

Fuel Cell Powered HEV Design and Control. Table 1 Different types of fuel cells

Parameters	PEM		DMFC	AFC	PAFC	MCFC	SOFC
	LT-PEM	HT-PEM					
Fuels	H ₂	H ₂ , Reformates	Methanol	KOH (H ₂)	PO ₄ (H ₂)	H ₂ , CO, CH ₄	H ₂ , CO, CH ₄
Charge carrier	H ⁺	H ⁺	H ⁺	OH ⁻	H ⁺	CO ₃ ²⁻	O ⁻
Electrolyte	PEM	PBI	PEM	30–50% conc. KOH in water	100% conc. in PO ₄	^a	^b
Temperature range	50–80°C	130–200°C	50–80°C	200°C	150–220°C	650°C	650–1,000°C

LT-PEM low-temperature PEM that are made of poly-sulphonated membranes, such as NAFION marketed by Du Pont, GORE membrane, ASAHI glass, 3M, etc.

HT-PEM High temperature membrane, such as polybenzimidazole (PBI) membranes made by BASF (USA), etc.

^aLiquid molten carbonate salts, such as Na₂CO₃ and Li₂CO₃, in a ceramic matrix of LiAlO₂

^bSolid oxide ceramic (Yttrium-stabilized zirconium (ZrO₂/Y₂O₃))

with nitrogen gas till it cools to ambient conditions to avoid leaching of the acid and degrading of the membrane performance [3].

Source of Fuels for Fuel Cell

The current earthquake in Japan and Virginia, USA, is a reminder that the nuclear power-generating plants are very vulnerable to natural disasters. Nuclear plant (prior to the earthquake in Japan) has been projected to be the main source of hydrogen production, as it is cleaner and more efficient. The Chernobyl (1986) nuclear disaster and the Japanese (2011) disaster triggered by an earthquake of 9.0 magnitude on the Richter Scale, clearly drew a line between science, engineering, and nature. Despite the precautions and technological advancement of Japan, it could not contain such magnitude of earthquake. Couple of months after the incident, the levels of radiation at the Japanese nuclear meltdown have been upgraded to level 7, same as the Chernobyl radiation after the disaster. The Virginia quake of summer 2011 rattled Washington D.C., metropolis with a magnitude of 5.8 on the Richter Scale. This spontaneously shutdown the nuclear reactor in Virginia, raising serious concerns of possible nuclear radiation to the atmosphere. These recent incidents pre-suggest that another sustainable source of hydrogen production (and not nuclear power plant) is needed for the hydrogen economy. Following the above Japanese event, Japan, Germany,

and Switzerland have frozen proliferation of building further new nuclear reactor plants.

The question now is which is the most environmentally and sustainable way of producing hydrogen for the hydrogen economy? Current pathways of hydrogen production include: steam methane reformation (SMR), partial oxidation (POX) process, and autothermal reformation (ATR) process of fossil fuels. However, the production and availability of fossil fuels are projected to dip after 2020, and as such, a renewable source of hydrogen is required. Electrolysis of water is a viable source; however, energy is required to electrolyze water and compress it to 350 bar (5,000 psi) or 700 bar (10, 000 psi) in order to store enough hydrogen to meet the same driving distance as its gasoline equivalent and package the fuel tank into the available fuel storage space. First, an analysis of the five types of fuel cells will be presented highlighting their potentials for hybrid vehicle propulsion.

In the final analysis, water electrolysis method to obtain hydrogen is a clean technology, but requires energy for the process and storage of the gas under pressure or possibly liquefaction of the fuel. The current technology of steam reformation relies on fossil fuels as feedstocks. Future energy carrier should be sustainable, clean, and renewable. Fuel cells for transportation in hybrid vehicles are feasible with the scalable fuel cell categories, such as PEM, direct methanol fuel cell (DMFC), alkaline fuel cell (AFC), and to a limited extent, with the solid oxide fuel cell (SOFC) as an auxiliary power unit (APU).

Fuel Cell as a Power System

Generally, fuel cells have the following advantages over other power-producing devices:

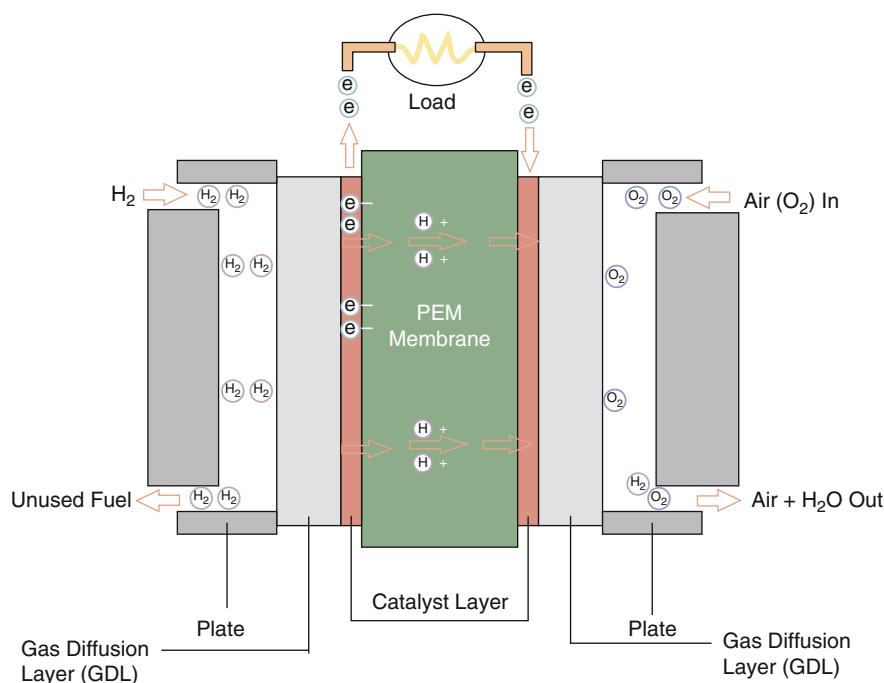
1. Low to zero emissions, if a reformat or hydrogen is used as fuel, which makes it environmentally very friendly.
2. Performance and efficiency are very high as there are no moving parts (excluding the system's balance of plant, BOP).
3. It is a bridge to hydrogen economy. The bridge may be in the form of electric vehicles (EV) or hybrid electric vehicles (HEV) – in which both fuel cell and a smaller internal combustion engine (ICE) are used for propulsion in the medium and low-power applications.
4. Absence of transmission lines which decreases the complexity of the BOP.
5. Flexible, scalable, and adaptable where remote power systems are required, for example, remote stations, booster stations, traffic and bacon lights, road signs, street lights, etc.

6. For space missions. Fuel cells have long been used for manned space missions, for example, the Apollo Mission of the 1960s and the Shuttle Orbiter.

In the following analysis, much attention is given to PEM because of its potential as the viable candidate fuel cell for use in transportation. It is noteworthy that not all types of fuel cells are suitable for use in the transportation industry because of the power density, type of electrolyte, size, and thermal cycling.

Analysis of Various Fuel Cells

The Proton Exchange Membrane (PEM) Fuel Cell The PEM fuel cell derives its name from the electrolyte used, which is a conducting membrane sandwiched between two catalyst layers, gas diffusion layers and plates. The schematic is shown in Fig. 1. A PEM fuel operates by inducing hydrogen into the anode, where it is oxidized into electrons and protons as per the equation:

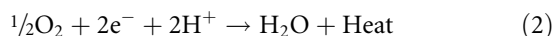


Fuel Cell Powered HEV Design and Control. Figure 1

PEM fuel cell components

The two electrons are not membrane permeable; hence, they travel through the anode electrical interconnect to the load, generating electricity. From there, the electrons travel to the cathode catalyst layer, where they react with the supplied oxygen and the protons that permeate through the membrane to the cathode catalyst layer to form water in a reduction reaction. The process is exothermic.

At the cathode, the reaction is as follows:



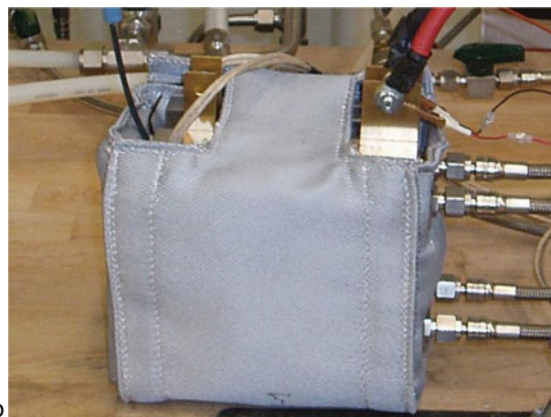
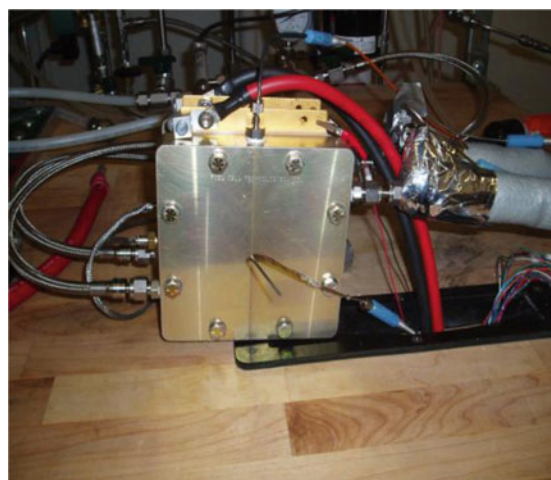
Overall reaction:



The schematics of a PEM fuel cell, a 50 cm² LT-PEM and HT-PEM fuel cell hardware are shown in Figs. 1, 2a and 2b respectively, while and multi-test station is shown in Fig. 3.

Figure 3 features a SERC (Shartz Energy Research Center) four test station (TS) platform for testing and validation of a single cell to 1 kW stack. The two TS to the left of the computer monitor operate in a dead-end mode and are capable of testing single cells and low-end stacks rated up to 250 W. While TS 3 and 4 are rated up to 1 kW and operate in a flow-through mode. Low-temperature PEM relies on humidification of the reactants and employs various types of humidifiers. For example, the 4-cell PEM stack on TS 2 has an-inbuilt membrane humidifier while TS 4 has an externally mounted humidifier shown in Fig. 4. That aside, PEM fuel cell stacks require a cooling system which also serves as the medium for maintaining the cell/stack temperature at a constant temperature. The test equipments are designed and sized for testing various output power ranges of cells/stacks. As the power output of the stack increases, the measuring devices ranges and pipe sizes are progressively increased to deliver the desired amount of reactants. Figures 5 and 6 show test equipment for low and high-temperature PEM fuel cell test equipments capable of testing and validating a 5 kW PEM stack.

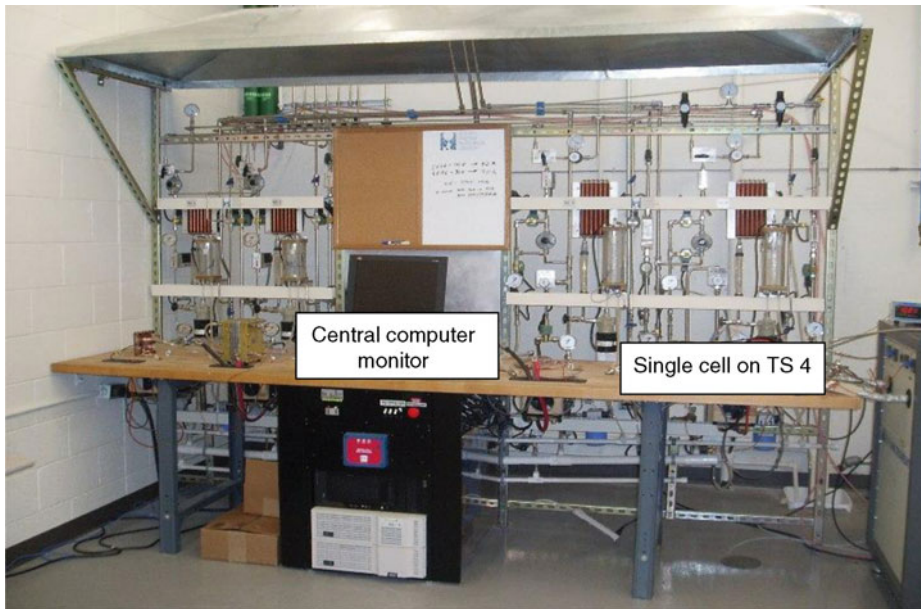
For the mobility sector, low-temperature PEM requires a complex balance of plant for its smooth operation. That aside, water from the exhaust is an issue of consideration during city driving in winter. It can contribute to slipperiness of the road and can



Fuel Cell Powered HEV Design and Control. Figure 2 (a) LT-PEM single cell of area 50 cm² assembled at the Center for Fuel Cell Research, Kettering University, Michigan (Courtesy Kettering PEM Fuel Cell Lab). (b) HT-PBI – PEM Single cell of area (Courtesy: Kettering PEM fuel cell research and powertrain integration, laboratory)

be a hazard to other road users. On the other hand, the high-temperature fuel cells offer an advantage in that vapor exits from the exhaust and can be used for heating the vehicle in winter. However, both low- and high-temperature PEM require heated pads to keep the stacks above 10°C during winter season. Freezing of liquid within the stack's flow field, in the gas diffusion layer and condensates in the exit flow manifold are all unacceptable.

PEM membranes are categorized as low-temperature and high-temperature membranes. A review of both



Fuel Cell Powered HEV Design and Control. Figure 3
Multi-test station (TS) for LT- and HT-PEM research



Fuel Cell Powered HEV Design and Control. Figure 4
Humidifier for a LT-PEM fuel cell at Kettering University Center for Fuel Cell Research

types highlights the following advantages and disadvantages: Both are credited with good startability – they start to generate electricity shortly after start-up (LT-PEM from 60°C and HT-PEM at 130°C). Performances of

PEM membranes are generally degraded by the presence of carbon monoxide. For LT-PEM, CO concentration above 10 ppm is not recommended, while HT-PEM based on a PBI can tolerate up to 3% CO in the fuel.



Fuel Cell Powered HEV Design and Control. Figure 5

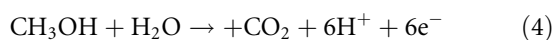
A 5 kW Greenlight low-temperature PEM test equipment at Kettering University Center for Fuel Cell Research

Higher concentrations have been tested up to 8% at a much reduced voltage [3–6]. Operation around 190°C favors higher CO concentration and the loss in performance is reversed if the cell is operated later with pure hydrogen [3].

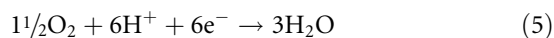
Both LT and HT-PEM require thermal management systems – LT-PEM uses water as the coolant while HT-PEM uses mineral oil. However, LT-PEM in addition relies on humidification of the membrane for proton transport. The functional humidification range is between 80% and 100% relative humidity (RH). Above 100% RH, the cell floods and impairs the flow of the reactants to the catalyst sites. As a consequence, the cell reaches its mass concentration limit, a point where the flow of reactants to the catalyst sites diminishes and the fuel cell voltage falls off. At a lower RH, the membrane dries up and impairs proton transfer from the anode to the cathode. This inadequacy of the LT-PEM led to the development of a more tolerant membrane, capable of operating without humidification and its complexities. Its electrochemical kinetics are enhanced because of its high-temperature operation; also, the activation overvoltage is reduced. It is more tolerant to a higher CO in the fuel composition.

Future fuels require flexibilities to fuel contaminants. Experience has shown that conventional fuels' dispensing, storage, and transportation systems (trucks, pipelines, etc.) harbor huge deposits of foreign materials that degrade the quality of fuel with time. High-temperature membranes are more resilient and less sensitive to increased CO constituents in the fuel than sulphonated membranes and can tolerate reformates. This enables reformers to be coupled directly to the fuel cell which eliminates the complexity of post-reformation processing of the fuel to obtain hydrogen of 99.997% purity.

Direct Methanol Fuel Cell (DMFC) The direct methanol fuel cell has the same architecture as hydrogen-air PEM fuel cell in that it uses sulphonated membranes. It is a low-temperature fuel cell that operates between 50°C and 80°C. Methanol is fed to the anode while air is fed to the cathode. The anode reaction proceeds as follows:



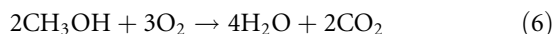
At the cathode, the reactions proceed as:





Fuel Cell Powered HEV Design and Control. Figure 6
A 5 kW high-temperature PEM fuel cell test stand at
Kettering University Center for Fuel Cell Research

The overall reaction is:

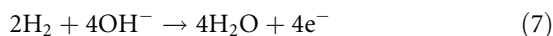


The reaction in DMFC is very slow in both electrodes despite its apparent advantages of releasing six electrons. To remedy this, a bi-metal catalyst Pt/Ru is used at the anode with a tenfold increase in catalyst loading compared with hydrogen-air PEM fuel cell [7–10]. The activation overvoltages in both electrodes of DMFC is considered, unlike hydrogen-air PEM where the anode activation overvoltage is ignored because it is by many orders of magnitude smaller than the cathodic activation overvoltage. DMFC is a candidate for portable devices such as cell phones, PDAs, soldier power, laptops, DVD players, iphones, ipads, ipods, kindles, and other portable devices. It works well in remote areas where power is required for a prolonged period such as remote sensing stations, railway crossings, traffic lights, bacon lights, professional cameras, etc. It is one candidate for light

HEV application but the technology has not been fully developed for such applications yet. DMFC has a low power density (kW/L); hence, its role as a power source for transportation is diminished.

Alkaline Fuel Cell The AFC uses liquid KOH electrolyte for its operation and it is popular for its past space missions in Apollo Spacecraft in the 1960s and later in the Space Shuttle Orbiter. Due to the renewed interest in the development of solid polymer membranes in the early 1990s, less attention has been given to AFC. The advantage of AFC is overwhelming. The electrolyte is cheap and has a proven technology for the production of KOH. However, the electrolyte is sensitive to the presence of atmospheric CO₂ in the air which combines with the potassium to form an insoluble K₂CO₃ that renders the electrolyte insoluble, hence degrading its performance. AFCs are installed in utility vehicles which operate within a restricted area where pollution by automobiles is restricted. It can be operated as a mobile and stationary power source.

The basic anode reaction



At the cathode



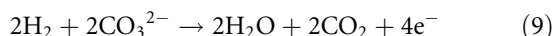
Phosphoric Acid Fuel Cell (PAFC) Phosphoric acid fuel cell PAFC derives its name from the type of electrolyte used. It is a medium temperature fuel cell used for stationary power generation. The operational temperature is in the 200–220°C range. It is one of the oldest and matured fuel cells with vast production in the ranges of 200 kW. However, more powerful units ranging in MW have been introduced into service. It finds wide applications as back-up power units in credit card companies' headquarters, police precincts, hospitals, banks, airports, etc. The USA and Japan are the two leading countries in the production and application of PAFCs. Phosphoric acid fuel cell is equally susceptible to CO poisoning as hydrogen-air fuel cell, but can tolerate up to 0.5% or 5,000 ppm of CO [4–7].

Molten Carbonate Fuel Cell (MCFC) Molten carbonate fuel cell is a stationary power generation system that operates around 650°C. It has the advantage of not using noble metals at catalyst as a result of high operational

temperature. It can use multiple fuels, such as methane (CH_4), CO, and hydrogen, as fuel. One advantage of MCFC as well as SOFC is the use of exit gases for cogeneration. That implies using waste heat to operate either steam or gas turbine, thereby bottoming the cycle and increasing the efficiency of the utilized fuel in the entire system. At the exit temperature of 600–650°C, the waste heat is rerouted to the boiler to generate steam to operate a steam turbine. The turbine in turn is connected to a generator or it can be geared to operate several devices of the balance of plant. The final waste heat at a lower enthalpy can be used for heating a facility. This is achieved through routing the city water and the waste heat through a heat exchanger, thereby reducing or eliminating the extra cost of maintaining the facility warm water system.

The common electrolytes used in MCFC are the carbonate salts of sodium, lithium, and potassium. The mobile ion is the CO_3^{2-} that moves from the cathode to the anode [7–10].

The anode reaction using hydrogen as fuel is:

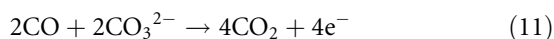


The cathode reaction is:



The reaction using CO as fuel is as follows:

At the anode, it is:



The cathode reaction is:

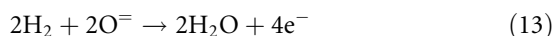


Molten carbonate fuel cell finds a wide application as a stationary power utility plant and is not suitable for hybrid electric vehicles purposes.

The Solid Oxide Fuel Cell (SOFC) The solid oxide fuel cell can be used both as a stationary and auxiliary (for mobile) power unit (APU) depending on its application. Current research favors using it as an auxiliary power unit for inter-state trucks, recreational vehicles, auxiliary power unit, and recharging unit for soldier power, power tools, etc. SOFC uses yttrium-stabilized zinc oxide as the electrolyte [7–10]. The operating temperature is between 650°C and 1,000°C. Due to its high operating temperature, SOFC as of now is not considered a candidate for hybrid electric vehicle

applications, but probably for future locomotive or metro-rail transportation systems.

The reaction at the anode using hydrogen as fuel is:

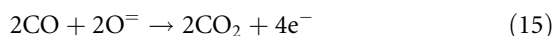


The cathode reaction is:



The reaction using CO as fuel.

At the anode, the reaction is:



The cathode reaction is:



Candidate Fuel Cells for Hybridization

The above overview shows three fuel cells as candidates for HEV: PEM, DMFC, and AFC in a descending order. Out of these three, AFC is least considered for automotive applications for the following obvious reasons: (1) The electrolyte is easily degraded in the presence of atmospheric CO_2 . When potassium carbonate is formed, less OH^- ions are available for ionization, and as such, the performance of the cell is degraded. (2) The fuel cell lacks flexibility in all possible orientations. The advantages are overwhelming: (1) The electrolyte is cheap and easily replaceable. (2) The technology for KOH production is available which makes the fuel source to be sustainable. (3) The absence of bipolar plates reduces the cost and technological process of AFC production [4].

Solid oxide fuel cell SOFC on the other hand can be considered in an auxiliary capacity to support recreation vehicles (RV), large trucks as an APU, but not as the main power unit. The disadvantages are: It requires some hours to bring it up to the operating temperature and shutdown – thermal cycling, and it operates at very high temperatures of 650–1,000°C and, as such, is not suitable for HEV application.

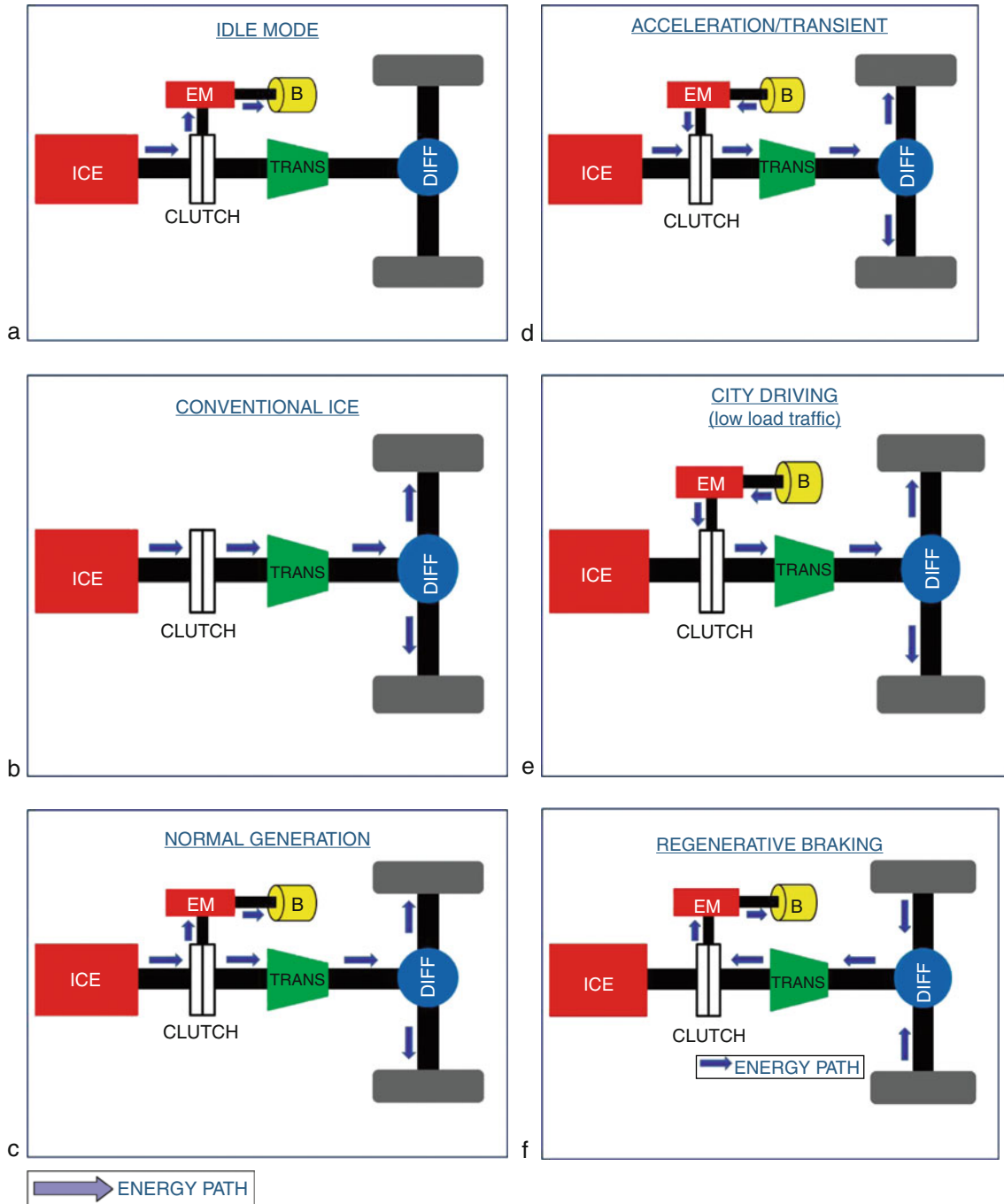
Hybrid Technology Using Gasoline or Diesel, Fuel Cells, Ultra-Capacitors, and Batteries

Hybrid Using Gasoline- or Diesel-Powered Engines

The technology of gasoline and diesel (internal combustion engine) operation is very well understood;

hence, little time will be devoted here for its explanation. Internal combustion engines (ICE) are devices that convert the chemical energy of the fuel (gasoline or diesel) to mechanical work through the piston,

connecting rod, and crankshaft mechanisms. The architecture of the gasoline or diesel hybrid consists of an ICE/generator, electric motor, battery, drivetrain, and wheels. The schematic is shown below in Fig. 7

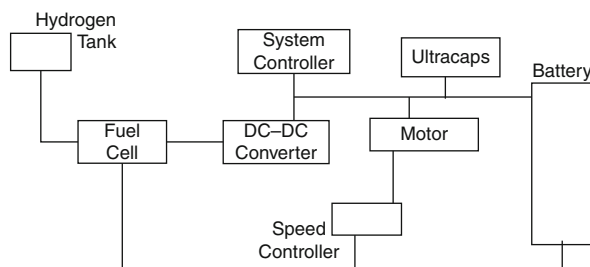


Fuel Cell Powered HEV Design and Control. Figure 7
Hybrid ICE

in which ICE/generator for oversimplification is represented as ICE. The mechanical power from ICE is converted into electrical power through the electric generator connected on the same shaft with the engine. Electric power is transmitted to the motor. Turbo-charged ICE has an immense advantage over its naturally aspirated versions. The exhaust gas energy from the turbine is harnessed by coupling an electric motor generator to the same shaft as the turbo-compressor taking advantage of the exhaust gas energy to generate extra electricity. During acceleration when the vehicle requires more torque/air, the electrical motor is activated with the extra energy from the turbo-generator. When there is excess exhaust energy, that portion would be harnessed to recharge the batteries, thereby bottoming the cycles of the exhaust gas use in hybrid vehicle [11]. The architecture in Fig. 7 shows various modes of ICE hybrid propulsion. Figure 7a presents idle operation where the generated power is used to charge the battery while (b) depicts the normal operation of a conventional ICE. The energy path is indicated with the arrows. Under normal operation, the power from the engine is used both for propulsion and charging the battery as shown in (c) while additional energy is needed during acceleration and transients from the second power source (battery) shown in (d). As an extender, the ICE is turned off during traffic in city driving and the energy from the battery is used for propulsion (e). The concept of regenerative braking involves capturing the kinetic energy during braking and deceleration to charge the batteries (f).

Fuel Cell Hybrid Architecture

The fuel cell hybrid vehicle architecture incorporates PEM fuel cell, onboard compressed hydrogen tank, and power conditioner, batteries, ultra-capacitors, motor and control electronics. A schematic of the PEMFC-powered hybrid vehicle using heavy-duty batteries and ultra-capacitors is presented in Fig. 8. The PEMFC balance of plant includes compressed hydrogen in a storage tank at a pressure of 350 bar (5,000 psi) or 700 bar (10,000 psi); cooling system; humidification system; compressed air system (fan, etc.); and an electronic display panel for controlling the flows into the fuel cell. The steps involved in integrating the entire



Fuel Cell Powered HEV Design and Control. Figure 8
Fuel cell hybrid architecture

system include sizing the fuel cell to provide average power over the entire operating range [12, 13]. The fuel cell plant has input and output data logging inlets and outlets to monitor the reactants' flow rates, the voltage, current, power, power density, and stoichiometries of reactants. A DC–DC converter is required between the FC and the load. Sets of batteries and ultra-capacitors are also required [12]. Charging of the batteries is done when the vehicle idles, or prior to using the vehicle. The DC–DC converter provides supplementary power for the ultra-capacitor's recharging functions while the ultra-capacitors provide power needs during transients, such as peak loads and transients during acceleration in fuel cell–powered vehicles. Some of the roadblocks in quick implementation of FC-HBVs are hydrogen infrastructure, sub-zero temperature operation, cold start constraints in the colder regions of the country and issues of dealing with fuel cell exhaust water during winter months on the road during city driving.

Ultra-Capacitors

The factor limiting the range and effectiveness of a rechargeable-battery-powered electric vehicle is the battery (or the energy storage system). Ultra-capacitors represent another generation of energy storage devices and are electrochemical double layer capacitors that supply peak energy requirement during transients, such as acceleration and cold start in fuel cell–powered vehicles and other hybrid electric vehicles (HEV). Figure 9 shows Maxwell's 2,600 F ultra-capacitors. Ultra-capacitors are widely accepted in fuel cell vehicles and other HEVs for their quick charge/discharge functions. Some of the advantages are their flexibilities in



Fuel Cell Powered HEV Design and Control. Figure 9
Maxwell ultra-capacitors

charging/recharging. They cannot be overcharged and can be maintained at any voltage above/below the rated value. While rechargeable batteries generally exhibit a very slow charge/response functions with significantly reduced life cycles, ultra-capacitors work best in this condition by recuperating within a few seconds as a result of their low equivalent series resistance (ESR). They possess a high cycle-life with less deterioration when used frequently and infrequently. They respond within a wide temperature range of 65–40°C. They exhibit high coulombic efficiency which implies total charge removed over the replenished charge [13–15].

Conclusion

In conclusion, both low- and high-temperature proton exchange membrane fuel cells will be used in fuel cell hybrid vehicles for transportation as they provide overwhelming advantages over the internal combustion engines, such as zero emissions when hydrogen is used as fuel, reduced noise, and higher efficiency.

The use of integrated technologies such as fuel cell, batteries, ultra-capacitors, and other novel techniques will occupy the center stage in fuel cell vehicle development within this decade.

Bibliography

1. Heywood JB (1985) Fundamentals of internal combustion engines. McGraw Hill, New York
2. Ferguson C, Kirkpatrick A (2001) Internal combustion engines and applied thermosciences. Wiley, New York
3. Ubong EU, Shi Z, Wang X (2009) Three dimensioning modeling and experimental study of a high temperature PBI based PEM fuel cell. *J Electrochem Soc* 156(10):B1276–B1282
4. Ubong EU, Dimitrov B (2010) Regression of the response variable of a high temperature PEMFC-PBI based membrane. *J Electrochem Soc* 157(7):B1059–B1067
5. Ubong EU (2008) A single cell testing and evaluation of a high temperature PBI based membrane. In: Fuel cell seminar & exposition. Convention Center, Phoenix, 27–30 Oct 2008
6. Ubong E, Das SK, Berry KJ, Antonio R (2008) Experimental performance evaluation of a high temperature PEM fuel cell at different temperatures. In: ASME 6th international fuel cell conference. Denver, 16–18 June 2008
7. Larminie J, Dicks A (2003) Fuel cell systems explained, 2nd edn. Wiley, New York. ISBN 0-470- 84857-X
8. Babir F (2005) PEM fuel cells: theory and practice. Elsevier, Amsterdam/London
9. Mench MM (2008) Fuel cell engines. Wiley, Hoboken
10. O'Hare R, Cha SW, Colella W, Prinz FB (2006) Fuel cell fundamentals. Wiley, New York
11. Sahed SM (2005) The power of turbocharging. *Automotive Engineering, International*, Sep 2005, SAE
12. Anzicek J, Joel Berry K, Lerner S, Slota G, Thompson M, Ubong E et al (2004) Fuel cell hybrid vehicle at Kettering University.

In: Electrical manufacturing and coil winding conference (EMCWA conference), Indianapolis, 23–25 Sep 2004

13. Ubong EU, Mizell C, Slota G (2005) Ultracapacitors with Ballard Nexa in a GEM vehicle in hybrid mode. In: Electrical manufacturing and coil winding conference) – EMCWA conference, Indianapolis, 20–22 Sep 2005
14. Schneuwly A, Maher B, Auer J. Maxwell Technologies Inc., www.maxwell.com
15. Namisnyk AM, Zhu JG. A survey of electrochemical supercapacitors technology. Faculty of Engineering, University of Technology, Sydney

Fuel Cell Types and Their Electrochemistry

GÜNTHER G. SCHERER

Paul Scherrer Institut, Electrochemistry Laboratory,
Villigen, Switzerland

Article Outline

Glossary

Definition of the Subject

Introduction

What Is a Fuel Cell?

The Role of the Electrolyte

Classification of Fuel Cell Types

The Role of Electrocatalysis and the Electrode/
Electrolyte Interface

Fuel Cell Families

Future Directions

Bibliography

Glossary

AFC Alkaline fuel cell.

EC Electrocatalysis.

FC Fuel cell.

GDE Gas diffusion electrode.

GDL Gas diffusion layer.

HOR Hydrogen oxidation reaction.

HT-PEFC High temperature polymer electrolyte fuel cell.

MCFC Molten carbonate fuel cell.

ORR Oxygen reduction reaction.

PAFC Phosphoric acid fuel cell.

PEFC Polymer electrolyte fuel cell.

SOFC Solid oxide fuel cell.

Definition of the Subject

Fuel cells are efficient energy converters, based on electrochemical principles. They convert the chemical energy (heating value) of a fuel directly into electricity, circumventing the various steps of thermal conversion and electricity generation. Fuel cells can be designed and constructed on the basis of a multitude of material combinations for electrolyte and electrodes, opening the choice of different fuels. The electrocatalytic reactions of fuel and oxygen are major challenges to obtain high conversion efficiency. The electrochemical basics of different fuel cell types considered today for technical applications are described in this contribution.

Introduction

Fuel cell technologies have seen a revival in recent years, due to several reasons. Global warming and local air pollution caused by various energy utilization processes have created a multitude of environmental concerns, promoting the development of novel technologies with high conversion efficiencies and low emissions, possibly zero emission, with respect to green house gases and other pollutants [1, 2]. Peak oil is another reason for the renewed interest in fuel cell technologies, in particular for automotive applications [3, 4]. Although this fact is discussed in a highly controversial manner, limitation in crude oil supply is obvious in the long-term perspective. This particular aspect of fossil fuel resources is strongly interlinked to the future perspective of the “oil price” and, hence, its economic competitiveness to other fuels, e.g., fuels from renewable sources. In this context, the installation of new supply infrastructures for alternative fuels, e.g., H₂, is an important additional economical and political factor. Dedicated analysis has clearly shown that energy conversion in fuel cells has to be based on fuels, in particular hydrogen, derived from renewable sources [5]. Further, the geographical distribution of oil reserves causes concerns about the supply security in industrial centers around the world.

Overall, there exist several reasons to ask for novel efficient conversion technologies for mobility (electromobility) and combined heat and power systems (CHP) with independence on fossil fuels, in particular crude oil. Another area of interest in fuel cell technology is portable electric and electronic applications, where the argument of potentially higher energy density as compared to today's available battery technologies, hence, longer time of operation, is of prime interest [6].

Electricity generation via fuel cells is an idea, based on fundamental research carried out by Christian Friedrich Schönbein at the University of Basel in the early nineteenth century. Schönbein's basic investigations of the respective reaction of H_2 gas (oxidation) and O_2 gas (reduction) dissolved in the electrolyte at an electrode/electrolyte interface, representing the fundamental fuel cell reactions, led William Robert Grove to the design and construction of the first operating fuel cell. He utilized H_2 as fuel and O_2 as oxidant at platinum metal electrodes in two separated closed compartments, filled with the respective reactant (hydrogen electrode, oxygen electrode), and joined by sulfuric acid as common electrolyte [7].

As will be shown in this chapter, the basic electrochemistry of a fuel cell, an electrochemical gas battery, is determined by the choice of the electrolyte and the two different electrodes immersed therein. In the course of time since this fundamental work of Schönbein and Grove, many combinations of electrolytes and electrode materials have been tested and fuel cell types derived upon there. This development has also been driven in the past by the choice of available fuel and possible application in view, in particular by space and military applications.

This introductory chapter will give an overview on the basic electrochemistry of some fuel cell types, developed today for dedicated technological applications. The respective electrochemistry will depend on the materials composition of the fuel cell, in particular, on the nature of the electrolyte and the temperature of operation. This further determines the choice of fuel, due to the strong influence of temperature on the mechanism of the respective electrocatalytic processes.

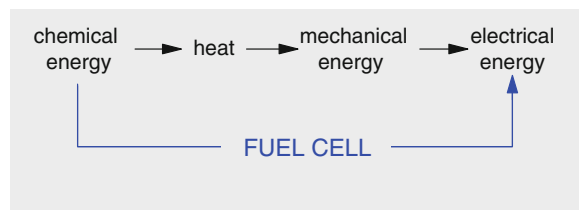
Engineering issues of fuel cell stack and systems design will be dealt with in the following chapters.

Hence, issues of, e.g., coflow or counterflow within one cell, stoichiometry and utilization of fuel and oxidant, temperature and current distribution in a fuel cell of technical scale, and, certainly, issues of stacking cells into a bipolar arrangement will not be discussed here.

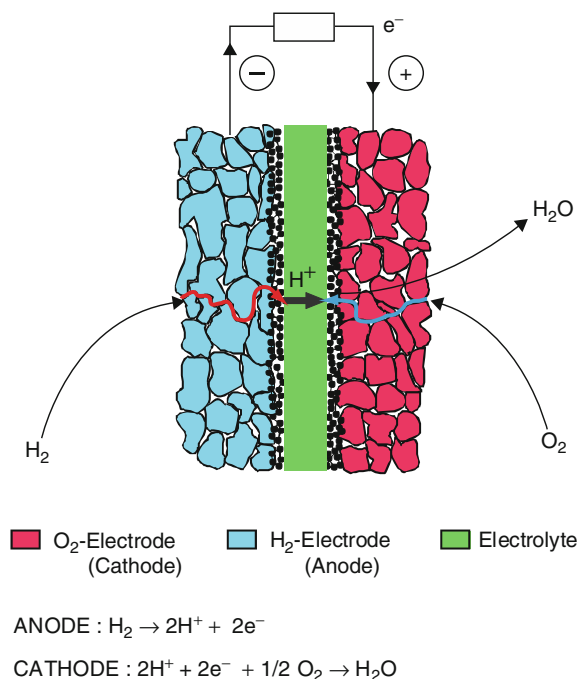
What Is a Fuel Cell?

A fuel cell is an electrochemical device, converting the chemical energy (Gibbs free energy) stored in a gaseous or liquid fuel, e.g., hydrogen, methane, methanol, ethanol, others, directly into *work of electrical energy (direct current electricity)* at constant temperature (Fig. 1). This type of energy conversion process is different from the classical thermomechanical energy conversion process and is not limited by the Carnot principle (see below).

In short, in a fuel cell, the fuel is oxidized at an electrochemical interface (electrode called anode), accepting electrons and donating these electrons at a second electrochemical interface (electrode called cathode, separated from the anode) to an oxidant, e.g., oxygen, which is reduced by accepting these electrons. Both electrochemical interfaces have to belong to a common electrochemical cell and are joined in the cell by a common medium, an ion-conducting electrolyte. Both electrodes have to be connected electronically by an external circuit, containing the electrical device to be operated, in which the electrons, due to the potential difference created by the two electrode reactions, travel from the anode to the cathode delivering electrical work (Fig. 2). Fuel and oxidant are supplied in gas channels of the cell housing (bipolar plate in stacked cells) on the backside of the porous electrodes (not



Fuel Cell Types and Their Electrochemistry. Figure 1 Scheme of direct energy conversion in an electrochemical cell as compared to "conventional" thermal conversion



Fuel Cell Types and Their Electrochemistry. Figure 2 Simplified fuel cell scheme with an acidic aqueous electrolyte, e.g., the polymer electrolyte fuel cell (PEFC). Fuel: H_2 , oxidant: O_2 . Only porous gas diffusion electrodes and electrolyte are shown, cell housing is not shown [8]

displayed in Fig. 2) Both gases have to be transported through the porous gas diffusion layers (GDLs) with pores typically in the micrometer range (blue and red bodies in Fig. 2) to the electroactive catalyst layers (ECLs, black dots in Fig. 2) at the interface to the electrolyte. Colloquially, GDL and ECL together are called gas diffusion electrode (GDE).

The fuel cell and its electrochemically active components, i.e., electrodes, electrolyte, etc., as well as its (electrochemically inert) structure materials, i.e., current collectors, cell housing, etc., should be as invariant as possible, i.e., they should not be consumed and, ideally, not age (corrode) over the time of operation. Hence, as an electrochemical reactor, they provide the electrochemically active interfaces (or interphases, see below) and the necessary pathways for mass transport for educts and products to and from these active interfaces through porous media (active electrode layers, gas diffusion layers, internally corrugated cell housing (flow fields) in bipolar plates) with open porosity at

different scales. At the same time, it is a prerequisite that these materials are as conductive as possible because they are responsible for the collection and transmission of the electric current generated at the two interfaces. Hence, ohmic voltage losses in these materials should be as low as possible.

The electrochemistry of the two electrode reactions is exemplified for the simplest and predominant case by the “cold” electrochemical combustion of H_2 with O_2 (pure O_2 or from ambient air) to H_2O (Fig. 2). The overall reaction is split into two partial reactions, occurring at the two different electrodes of the cell:

Anodic reaction : $\text{H}_2 = 2\text{H} + 2\text{e}^-$
(hydrogen oxidation reaction, HOR)

Cathodic reaction : $1/2\text{O}_2 + 2\text{H}^+ + 2\text{e}^- = \text{H}_2\text{O}$
(oxygen reduction reaction, ORR)

Overall reaction : $\text{H}_2 + 1/2\text{O}_2 = \text{H}_2\text{O}$

The reaction product(s) of the individual electrode reactions are constituents of the (aqueous) electrolyte (H^+ in the example of an acidic electrolyte, as displayed in the scheme of Fig. 2), which avoids polarization losses in the electrolyte (see Chapter “► Livestock Somatic Cell Nuclear Transfer”).

Each of the two electrode reactions creates a characteristic potential difference across the interface solid electrode/electrolyte, which is different for the two reactions according to the different reactants. The overall cell voltage between the two electrodes, which are joined by the same electrolyte, allows the electrons generated at the anode (HOR) and consumed at the cathode (ORR) to create work in the external circuit. Hence, chemical energy released by the individual electrode reactions at the locally separated electrodes is directly transferred into electrical energy. This pathway is different from the combustion step in the “classical” thermomechanical power generation, where the oxidation of fuel and reduction of oxidant occur in the same volume element, thereby generating heat only.

Generally, the available cell voltage of electrochemical cells depends on the thermodynamics of the two electrode reactions in the prevailing electrolyte, hence the difference in the electrode potentials, and is confined, according to the series of electrochemical potentials, to a few volts [9]. According to the individual

electrode potentials of the H_2/H^+ reaction (by IUPAC standard zero volt in the series of electrochemical potentials, acidic electrolyte, standard conditions of 1 atm and 25 °C or 298K) and the $\text{O}_2/\text{H}_2\text{O}$ reaction (1.23 V, respectively), a cell with H_2 and O_2 as reactants should yield an ideal cell voltage of 1.23 V at these standard conditions. In practice, a lower value in the range of 1 V is observed, due to different implications (side reactions, depolarization of electrodes due to crossover of gases through the electrolyte, etc.).

The free energy ΔG (Gibbs energy at constant pressure) of the fuel cell reaction is related to the cell voltage under open circuit conditions (open circuit voltage, OCV) E^0 and standard conditions according to

$$E^0 = -\Delta G/nF$$

n being the number of electrons transferred in the respective reaction (equal to 2 for the reaction $\text{H}_2 + \frac{1}{2}\text{O}_2 = \text{H}_2\text{O}$) and F the Faraday constant equal to 96,487 C/mol (electric charge per mol of electrons).

Further, ΔG can be expressed as the difference of the free enthalpy ΔH of the reaction (thermal cell voltage E_{H}^0 equivalent to 1.48 V for the H_2/O_2 reaction under standard state conditions) and the term $T\Delta S$, (T absolute temperature) expressing entropy losses (or gains) ΔS in the reaction:

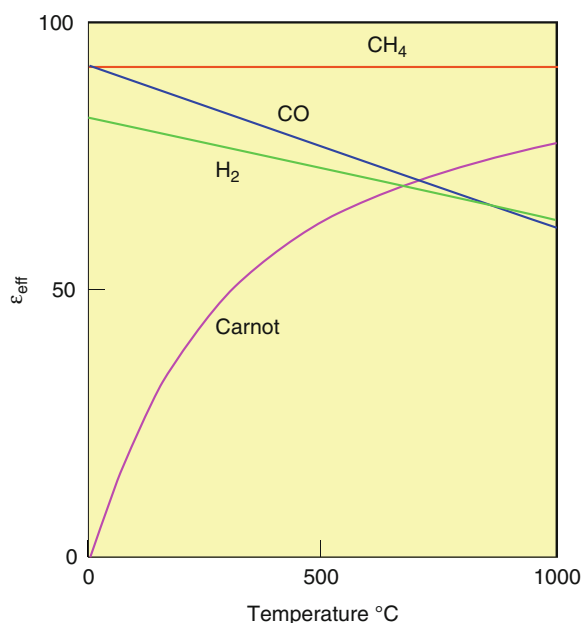
$$\Delta G = \Delta H - T\Delta S$$

Hence, the voltage (thermal) efficiency of a reaction under equilibrium conditions can be written as:

$$\varepsilon_{\text{th}} = \frac{\Delta G}{\Delta H} = 1 - \frac{T\Delta S}{\Delta H}$$

The voltage (thermal) efficiency for the electrochemical conversion of the three different fuels H_2 , CH_4 , and CO reacting with air (oxygen) in dependence of reaction temperature is displayed in Fig. 3. In comparison, also the maximal efficiency $\eta = (T_1 - T_2)/T_1$ for a Carnot process at the respective higher temperature T_1 and the lower temperature T_2 equal to 0 °C is included. The different temperature dependence of the three conversion reactions is due to the respective entropy changes, e.g., $\Delta S = 0$ for the reaction $\text{CH}_4 + 2\text{O}_2 = \text{CO}_2 + 2\text{H}_2\text{O}$ and $\Delta S \neq 0$ for the reactions $2\text{H}_2 + \text{O}_2 = 2\text{H}_2\text{O}$ and $2\text{CO} + \text{O}_2 = 2\text{CO}_2$.

Based on thermodynamic arguments only, electrochemical energy conversion offers higher efficiencies



Fuel Cell Types and Their Electrochemistry. Figure 3 Voltage (thermal) efficiency ε_{th} for the electrochemical reaction of H_2 , CO , and CH_4 as fuel with oxygen from air under equilibrium versus temperature (°C). In comparison, Carnot efficiency of a thermal process as function of the respective higher temperature in °C and with 0 °C as lower temperature is displayed

within a certain temperature window, depending on the respective reaction.

Introducing Nernst's law for the equilibrium case, the situation when no current (and hence power) is delivered by the cell, the equilibrium cell voltage under nonstandard conditions for a H_2/O_2 cell in dependence of the respective reactant/product concentration (partial pressures) can be expressed as:

$$\begin{aligned} E^{\text{eq}} &= -\frac{\Delta G}{nF} = E^{\text{eq,c}} - E^{\text{eq,a}} \\ &= E^0 + \frac{RT}{2F} \ln \left(\frac{[\text{H}_2][\text{O}_2]^{1/2}}{[\text{H}_2\text{O}]} \right) \end{aligned}$$

E^{eq} = equilibrium cell potential, $E^{\text{eq,c}}$ = equilibrium potential cathode, $E^{\text{eq,a}}$ = equilibrium potential anode, R gas constant, T absolute temperature, E^0 = equilibrium potential under standard state.

E^{eq} is also called "theoretical" open circuit cell voltage, OCV, in dependence of temperature and reactant,

respectively product concentrations. Practical OCVs may observed to be lower, due to losses (mixed potentials due to side reactions or gas crossover, see practical OCV of H_2/O_2 fuel cell) already occurring at zero current flow.

Under current flow, the kinetics of the electrode reactions comes into effect. The Butler-Volmer equation describes the dependency of the current passing through the electrode interface (current density per unit geometric area) on a small voltage excursion (called overvoltage η) from the respective equilibrium potential E^{eq} .

$$i = i_0 [\exp(\alpha n F \eta / RT) - \exp((1 - \alpha) n F \eta / RT)]$$

i = current density at overvoltage η , η = overvoltage (excursion from the equilibrium potential, i_0 = exchange current density (at $\eta = 0$), α = symmetry factor, representing the fraction of overvoltage assisting the reaction, n = number of electrons exchanged, F = Faraday constant, R = gas constant, T = absolute temperature.

The exponential dependency points out that small potential excursions cause a larger change in current, while at larger potential excursions, the current follows a linear behavior. This can be described by the Tafel law, which usually is applied when current densities of technical interest, e.g., several 100 mA per cm^2 geometric surface area in a fuel cell, are passing through the cell.

$$\eta = (RT/\alpha n F) \ln i/i_0$$

For the HOR, exhibiting an exchange current density in the range of $i_0 = 10^{-3} \text{ A/cm}^2$ with platinum as electrode material [9], only a small overvoltage arises even by drawing a current density of technical interest, e.g., 1 A/cm^2 , due to the fact that the H_2 molecule is easily split into two adsorbed H atoms (Tafel reaction), followed by the transfer of one electron per H atom (Volmer reaction) to the electrode generating two hydrated H^+ . Alternatively, one electron is transferred to an adsorbed H_2 molecule, followed by a split into a hydrated H^+ and an adsorbed H atom. In a consecutive step, the second H atom is oxidized (Heyrovsky reaction).

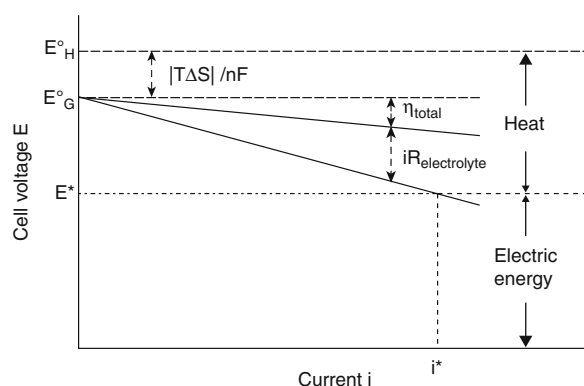
In contrast, the exchange current density of the ORR at a platinum electrode is in the range of $i_0 = 10^{-8} \text{ A/cm}^2$. Assuming *direct* reduction of the oxygen molecule, the ORR requires four electrons in

total, including the four H^+ generated in the HOR at the counterelectrode (see above). As a consequence, ORR generates a major overvoltage and, hence a major contribution to the overall loss.

In addition to the overvoltage losses due to electrode kinetics, one has to take into account the voltage loss in the ion-conducting electrolyte, due to the finite electrolyte conductivity, and (minimized) losses due to the electric resistance of electrode and cell materials, including contact resistances.

This relationship between cell voltage and current density is schematically displayed in Fig. 4. The various voltage losses are deducted from the equilibrium cell voltage (zero current), yielding the available cell voltage E^* at the respective current density i^* . The difference between E_{H}^0 and E^* is expressed as (waste) heat. In terms of voltage efficiency, this (heat) loss should be as low as possible; hence, E^* should be as high as possible. On the other side, the current density to be drawn has to be as high as possible to utilize power output values (specific power W/kg or power density W/L) of technical interest.

Further, Fig. 4 points to another advantage of (electrochemical) fuel cell technology: Designing a fuel cell to the maximum or rated power for a certain application (power = cell voltage E times current i), one easily realizes that the voltage efficiency ϵ_{eff} increases at lower power (lower current but higher voltage). This fact is



Fuel Cell Types and Their Electrochemistry. Figure 4 Cell voltage E versus current i of a fuel cell, with E_{H}^0 = thermal cell voltage, E_{G}^0 = ideal cell voltage based on ΔG , η = sum of overvoltage at anode and cathode, $R_{\text{Electrolyte}}$ = electrolyte resistance

particularly important for applications where maximum power (full load) is only required at certain parts of the duty cycle, while in other parts, only part load is necessary. This case is, e.g., exhibited in automotive applications of fuel cell technology.

Generally, the available cell voltage of electrochemical cells depends on the thermodynamics of the two electrode reactions in the prevailing electrolyte, hence the difference in the electrode potentials, and is confined, according to the Electrochemical Series of Standard Potentials, to a few volts. Cells with an aqueous electrolyte exhibit a limitation given by the stability window of water, namely 1.23 V at standard conditions. As stated above, the H_2/O_2 fuel cell allows practical open circuit voltages of around 1.0 V. At cell voltages above 1.23 V, typically around 1.5 V, decomposition of water into H_2 and O_2 occurs.

Hence, to accumulate the necessary voltage for technical applications, e.g., 200–400 V, for an electrical power train in a car, cells must be connected in series. Dedicated bipolar arrangements of cells have been designed and put into operation for serial connection, taking into consideration also the necessary parallel mass flow of fuel and oxidant from a manifold into each individual cell and the respective removal of the product. Such an arrangement of cells is called a fuel cell stack, combining the electrical serial connection of individual cells with a parallel connection for mass flow.

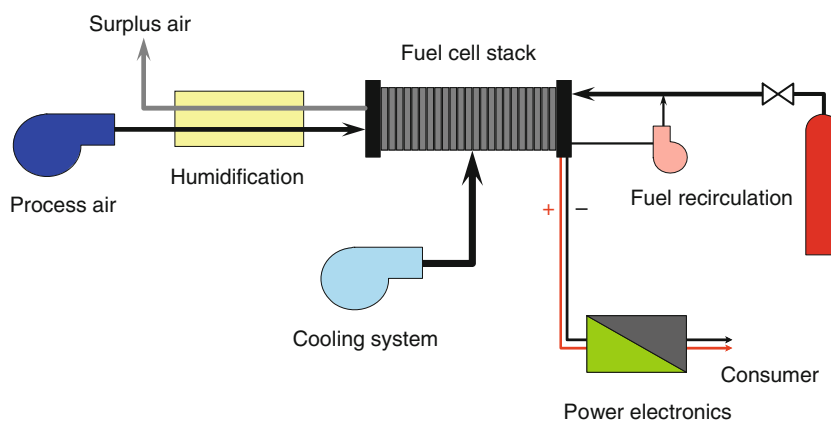
In contrast to batteries, fuel cells are open systems, which convert the chemical energy available in a fuel

stored outside the fuel cell, the electrochemical converter. As a consequence, fuel cells need a fuel tank, also a tank for the oxidant, if the oxidant is pure oxygen and not ambient air, and auxiliaries (for temperature, pressure, etc., control) to be operated (Fig. 5).

Mass flow into (reactants) and out of the cell (products) must be allowed and controlled, as well as cooling to remove the waste heat. At a voltage efficiency of 50%, a cell producing 1 kW electric power also generates 1 kW of heat, which has to be removed efficiently not to cause thermal runaway and, hence, degradation of the cell and its components. In summary, this is called a fuel cell system. Engineering issues of mass and heat transfer, current distribution across the active area of a cell, and voltage distribution along the stack, etc., are governing the design of a fuel cell system.

Fuel cells can be operated continuously, as long as they are supplied with fuel and oxidant. Hence, they are not limited by electrical charge–discharge cycles, as batteries are. This fact can be of utmost importance for certain applications, e.g., with respect to refueling time in mobile applications. For mobile or portable applications, the availability of a fuel infrastructure adds another aspect of consideration.

In terms of achievable specific energy (W/kg) and volumetric energy density (W/L), air-fed fuel cells have the advantage that one reaction partner, namely, oxygen, is derived from air, and consequently does not add to the weight and volume of the device. However, ambient air must be processed (cleaned, humidified, etc.), which has to be taken into account.



Fuel Cell Types and Their Electrochemistry. Figure 5

Simplified scheme of a fuel cell system (Adapted from L. Gubler, Paul Scherrer Institut)

This advantage is also expressed in the concept of metal air batteries, e.g., the Zn/Air or Al/Air battery, which on the anode (metal) consists as half a battery (closed mass flow) and on the cathode as half a fuel cell (open mass flow).

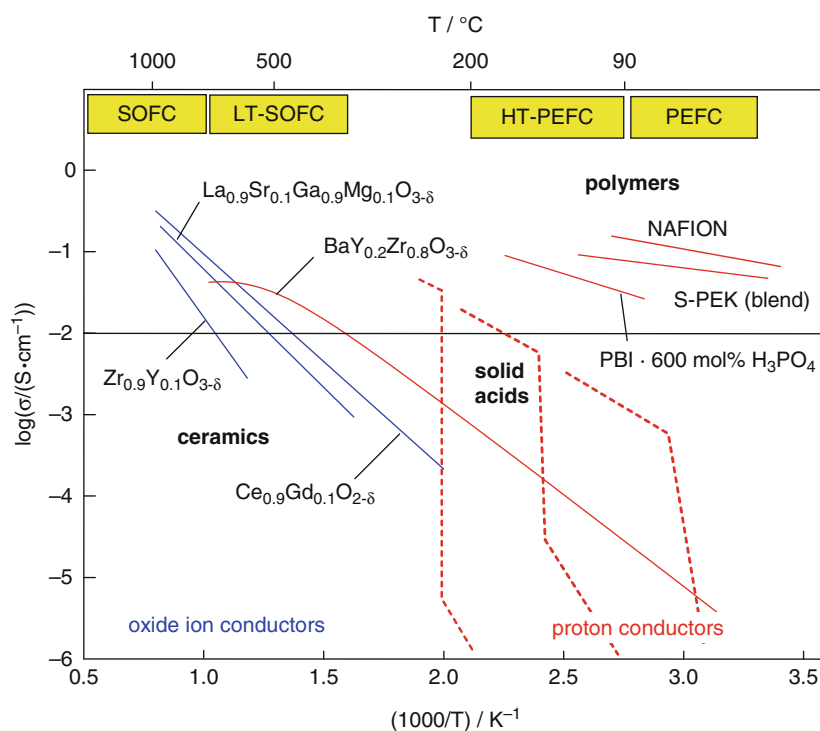
Next to the multitude of material combinations allowing various fuel cell types to be designed and built, the variety of mobile (transport), stationary, and portable applications provokes a multitude of different cell, stack, and systems designs, guided by the implications of the different fuel cell technologies (see below).

The Role of the Electrolyte

Generally speaking, the ion conduction in the electrolyte contributes to the losses (ohmic losses, see Fig. 3) in the cell voltage. Hence, one of the major tasks in fuel cell development is the reduction of these losses in the ionic circuit part, because the work available in the external electronic circuit should be as high as possible.

The role of the “electrolyte” in a fuel cell is a multiple one:

1. Transport the charge in ionic form from one interface to the other within the cell. To minimize losses in the ionic circuit, the ion conduction should be carried by an ion, which is produced at one electrode and consumed at the other to avoid losses caused by concentration gradients [10]. This majority ion should carry as much charge as possible through the electrolyte, hence its transport number (Hittorf number) should be as high as possible, ideally one. For high power applications with an aqueous electrolyte, the concentration of this respective anion or cation must be high and exclusively responsible for conduction. Typically, a specific ionic conductivity in the range of 100 mS/cm is required, which at a current density of 1 A/cm² and an electrolyte gap of 100 μm would yield an ohmic voltage loss in the range of 100 mV. As seen in Fig. 6, the conductivities of various



Fuel Cell Types and Their Electrochemistry. Figure 6

Specific conductivity versus temperature of different electrolytes interesting for fuel cell applications (Adapted from K.D. Kreuer)

electrolytes are strongly temperature dependent and cover a wide temperature range, depending on the ion-conducting species in its respective environment (material). Again, electrolytes are of interest, which offer the opportunity that the ionic species is participating in the fuel cell reaction.

2. Act as separator to (1) avoid crossover and, as a consequence, “hot” combustion of the fuel and the oxidant at the respective counterelectrode, and (2) to avoid touching of the electrodes by short circuiting the electronic pathway. Thereby, the electrolyte/separator gap should be as narrow as possible to lower the ohmic contribution to the overall voltage loss. In the case of a liquid electrolyte, it can be absorbed into an inert matrix, providing porosity high enough for an ionic conduction path with low tortuosity for the ion and at the same time a high enough bubble pressure to suppress gas crossover. An ideal concept is the one of a solid electrolyte, which fulfills the dual electrolyte and separator function at the same time, providing a transfer number of one for the current-carrying ionic species, as in the case of solid polymer electrolytes (ion exchange membranes), thereby excluding a contribution of electronic conduction.

One has to emphasize that the electrolyte has to sustain the potential window of the respective fuel cell reaction, given by the Gibbs free energy, at least over the device lifetime, specified for a certain application.

In aqueous acidic electrolytes, the ionic conduction is provided by an H^+ ion, respectively, $H(H_2O)_n^+$. As mentioned above, H^+ is created by the anodic oxidation of H_2 as a fuel, and the conducting species is transported to the cathode, where it is consumed in the direct (ideally four electrons) cathodic reduction of molecular oxygen. Hereby, water as the reaction product appears at the cathode side of the cell.

In alkaline solution, the ionic conduction is provided by OH^- ions, created by the ORR at the cathode. Principally, OH^- ions are carriers for the O^{2-} ions, produced as intermediate by the cathodic reduction reaction of molecular O_2 and the follow-up reaction with a water molecule. After conduction, OH^- reacts with the H^+ created at the anode and yields water as a product at the anode side.

Solid polymer electrolytes on the basis of a polymer cation exchange membrane in H^+ form or an anion exchange polymer membrane in OH^- form can be considered as quasi-aqueous electrolytes, whereby the water is absorbed in the phase separated ionic nanomorphology of the respective material. This nanomorphology forms ionic pathways through the polymeric membrane connecting the two fuel cell electrodes.

In cells operated at temperatures below 100 °C, liquid water will be the reaction product. For operation temperatures above the boiling point of water, a cell with an aqueous or quasi-aqueous electrolyte can be operated, however, at the expense of pressurizing it to avoid loss of the water and, as a consequence, concentration and conductivity changes in the electrolyte.

An alternative electrolyte for operating temperatures above 100 °C, which also gained technical relevance, is orthophosphoric acid, H_3PO_4 . “Water-free” orthophosphoric acid is a self-ionizing amphoteric system, which yields $H_4PO_4^+$ and $H_2PO_4^-$. Protons can be passed along pyrophosphate chains at elevated temperatures of around 200 °C, at which product water appears as vapor and can be rejected easily.

Molten salts can also be utilized as electrolytes. Again, the ion responsible for conductivity should be created in one electrode reaction and consumed in the other. However, due to the (1) corrosive nature, (2) the limited thermal stability, and/or (3) the high melting point of many melts, only a few molten salts have been considered. Up to now, the only fuel cell technology based on a molten salt electrolyte has been using a eutectic mixture of alkali metal carbonates, Li_2CO_3 and K_2CO_3 , at temperatures of around 650 °C, absorbed in a matrix of $LiAlO_2$. In this molten carbonate fuel cell (MCFC) the ion conducting the current is O^{2-} in the form of CO_3^{2-} . At the anode, CO_2 is supplied and reacts with O_2 and $4e^-$ to form CO_3^{2-} , which carries the current to the anode, where it is released in the hydrogen oxidation reaction. As a consequence, MCFC, in addition to the mass flow of fuel and oxidant, needs mass flow management of CO_2 .

At even higher temperatures in the range of 800–900 °C, ionic conduction can also be observed in solid oxides. Certain doped metal oxides become O^{2-} conductors; hence, they can be utilized as electrolytes in the solid oxide fuel cell (SOFC). The most

common oxide used is yttria-stabilized cubic zirconia (ZrO_2), called YSZ, with a doping level of 8% Y_2O_3 . This material is a ceramic; the thermal expansion coefficient of its thin layers has to be matched with the other fuel cell components.

As indicated by the terminology of fuel cells, e.g., PEFC, AFC, SOFC, others (see below), the electrolyte is the decisive cell component, which determines operation temperature, the choice the electrodes (electrocatalytic materials), and finally the specifics of the electrochemistry of the reactants. This, on the other hand, has consequences for the layout of the fuel cell design (balance of plant) and possible applications of the respective fuel cell technology, according to the required duty cycle of the application.

Classification of Fuel Cell Types

Today's fuel cell types under development can be classified in two different ways: either by the temperature range of operation or by the nature of the electrolyte.

The latter classification distinguishes between acidic and alkaline fuel cells. Acidic fuel cells with H^+ conduction require other electrode materials than alkaline with OH^- conduction. Alkaline fuel cells, at least with an aqueous electrolyte, require a CO_2 -free fuel, and hence cannot be operated on reformat fuel obtained originally from a C-containing fuel.

The above-described situation displays the fact that there is a temperature gap in the availability of electrolytes between ca. 200 °C and 650 °C. Today, major research efforts are carried out to extend the range of aqueous H^+ and OH^- conduction to temperatures beyond 100 °C, for technological reasons (heat rejection in cars, CO tolerance of platinum-based catalysts). On the other hand, oxide materials are developed with improved O_2^- conductivity at temperatures lower than the operation temperatures of standard SOFCs. Also ceramic proton conductors [11] and proton-conducting salts like CsHSO_4 [12] are considered as electrolytes for this temperature regime.

The Role of Electrocatalysis and the Electrode/Electrolyte Interface

Next to the voltage losses due to ionic conduction, voltage losses due to the activation overvoltages of the respective electrode reactions arise, as described by the

Butler-Volmer, respectively, the Tafel equation. As stated above, at a certain value of the current density, the overvoltage for the HOR is much smaller than for the ORR, due to the simpler electron transfer kinetics. Overvoltage for an electrochemical reaction shows also a strong temperature dependency, as does activation energy, and decreases with increasing temperature.

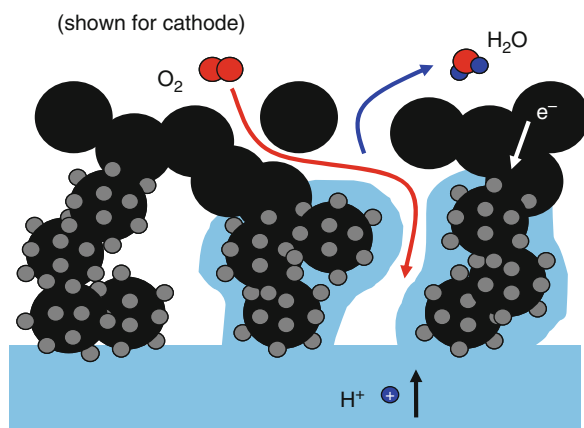
For fuel cells with an acidic electrolyte, platinum is the electrocatalyst of choice at temperatures up to 200 °C (e.g., PAFC), due to the stability requirements. High dispersion of the catalyst is required, taking into account its nature as a precious metal, to provide a high surface area to volume ratio. Platinum nanoparticles in the diameter range of a few nanometers, typically 2–5 nm, supported on carbon particles are utilized. Different carbons with different surface areas (up to 800 m^2/g) are used. The amount of platinum in the catalyst powder is specified as wt% per g of carbon and further as mg loading per cm^2 of geometric electrode area.

Platinum is prone to CO adsorption at temperatures of around 100 °C and below. For this reason, care has to be taken with respect to the purity of hydrogen, in particular for hydrogen liberated from a C-containing fuel by a reforming process (steam reforming, partial oxidation, autothermal reforming).

Alkaline electrolytes allow cheaper, non-precious metal catalysts like nickel and its alloys.

High temperature fuel cells, e.g., MCFC or SOFC, require other catalyst materials, which are described in the respective section below.

There is a common problem to all material selections for the different fuel cell types, namely, to generate an optimal electrolyte/electrode interface. One has to realize that gaseous molecular reactants are converted into ionic species, solvated/hydrated in their electrolyte medium, by exchanging electrons with the electrode material. This reaction occurs at the so-called triple phase boundary (triple point), where electron conducting, ion conducting, and gas phase (eventually dissolved in the electrolyte) join to each other. To allow a high surface area for current generation, the electrolyte/electrode interface is extended into a three-dimensional *interphase* with a certain thickness. For a polymer electrolyte fuel cell, this *interphase* is drafted in Fig. 7 for the cathode (oxygen) side.



Fuel Cell Types and Their Electrochemistry. Figure 7 Electrode layer (interphase) with three phase boundary (schematic) of a polymer electrolyte fuel cell (cathode side). Blue: polymer electrolyte, black: carbon particles, gray: platinum nanoparticles (Adapted from L. Gubler, Paul Scherrer Institut)

Depending on the respective materials for the different fuel cell types, this three-dimensional interphase has to be established by very different preparation/manufacturing procedures. In particular, solid electrolyte materials must be processable to be included into the electrode layer.

Fuel Cell Families

The various fuel cell types considered today having some technical relevance are displayed schematically in Fig. 8. They are arranged according to the nature of electrolyte, acidic or alkaline, and to their temperature of operation.

The polymer electrolyte fuel cell, PEFC, and the phosphoric acid fuel cell, PAFC, are acidic fuel cells; the PEFC operates in the temperature range below and around 100 °C (PEFC), respectively, and the PAFC at around 200 °C. Hydrogen is the preferred fuel for both types. A PEFC can also be fed by liquid or gaseous methanol, called direct methanol fuel cell, DMFC. Other fuels based on alcohols, e.g., the direct ethanol fuel cell, DEFC, are subject to research. Fuel cell types utilizing an acidic aqueous electrolyte will not be considered here.

Most concepts for an alkaline fuel cell in the past have been described with an aqueous alkaline electrolyte.

Recently, also anion exchange membranes, alkaline solid polymer electrolytes, have been considered [13].

The molten carbonate fuel cell, MCFC, and the solid oxide fuel cell, SOFC, are high temperature fuel cells, due to their temperatures of operation of around 650 °C, respectively, 800–900 °C. Both can be considered as alkaline fuel cells.

The local appearance of the reaction product is determined by the current-carrying ion. While for acidic fuel cells the reaction product shows up at the cathode (oxidant side), it is the anode (fuel side) for alkaline fuel cells. In case of C-containing fuels, e.g., methanol, a second reaction product containing the oxidized carbon appears at the fuel side. The appearance and hence the removal of the product/s causes consequences for the mass flow engineering of the stack and fuel cell system.

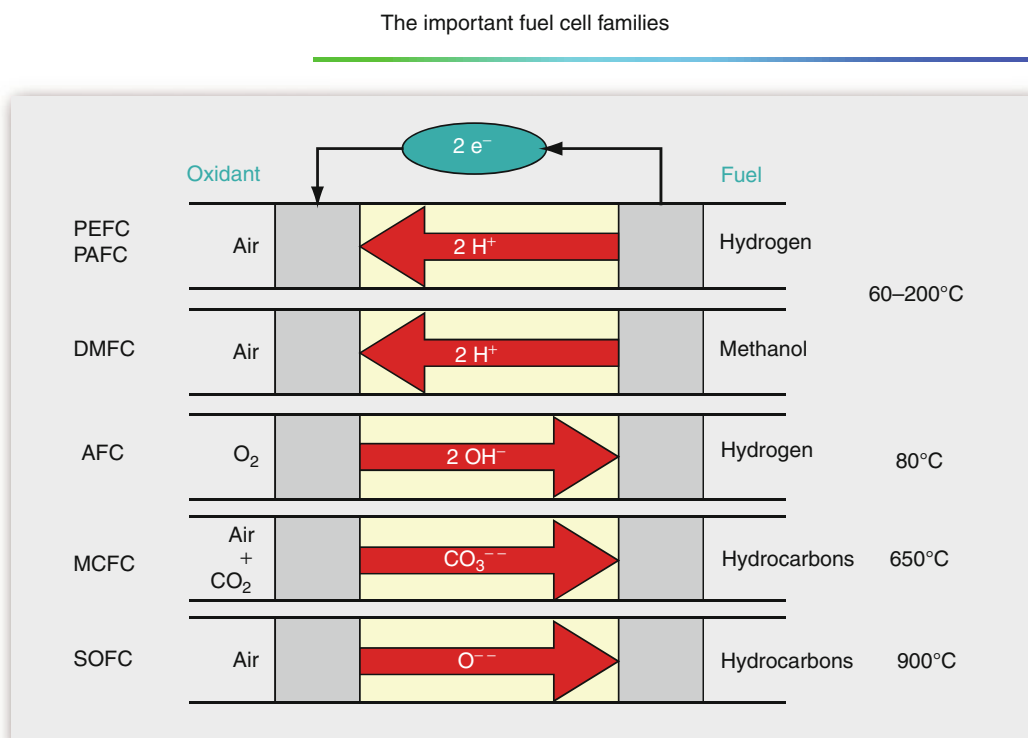
The following ordering of fuel cell types does not reflect the historical development of the various fuel cell technologies.

Polymer Electrolyte Fuel Cell Types

The idea of using a thin ion-conducting polymer membrane, *solid polymer electrolyte*, as electrolyte and separator can lead to different concepts of cells. Firstly, when ionic charges, anions or cations, are chemically bound to the polymer (ionomer) network, the respective countercharge can move freely within the polymer volume, provided a certain volumetric charge density within the polymer exists, which on uptake of, e.g., water leads to phase separation into a hydrophobic polymer (backbone) and a hydrophilic, charge-containing phase [14]. As a consequence, the polymer morphology allows the continuous transport of this ion from one electrode interphase to the other. This concept is normally called polymer electrolyte fuel cell (PEFC).

Depending on the fuel, H₂, MeOH (gas or liquid), or others, the solid polymer electrolyte properties have to be tailored, in particular for their properties toward separation of fuel and oxidant.

Further, non-ionomeric polymers can absorb an electrolyte, e.g., phosphoric acid H₃PO₄, which then renders ion conducting. This concept is realized in the so-called HT-PEFC, high temperature PEFC (ca. 100–200 °C), where a polybenzimidazole film



Fuel Cell Types and Their Electrochemistry. Figure 8

The important fuel cell families of today

imbibed with phosphoric acid becomes proton conducting due to the mechanism described above for non-aqueous phosphoric acid (see this book, Chapter Benicewicz).

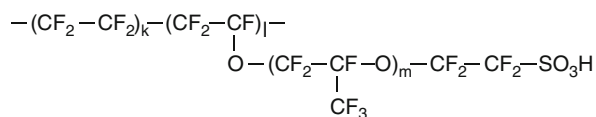
H₂-Fed PEFC This type of H₂-fed fuel cell, due to its high achievable specific and volumetric power density (in the range of 1 kW/kg and 1 kW/L and above), its cold start behavior, and its fast load following properties, has found interest in the automotive industry and is under development by all major automotive companies.

A thin ion-conducting polymer sheet, typically in the range of 25 μm , is utilized as a solid electrolyte, i.e., electrolyte and separator, in the PEFC (Fig. 1). For thermal stability reasons, only cation exchange membranes in the H^+ form have been considered up to today for technical applications (see also “Alkaline Fuel Cell”). In comparison to cells with liquid electrolyte, this PEFC concept offers the advantage that the

“electrolyte” is chemically bound within the polymer matrix and only water as reactant, in addition to the gases, appears in the peripheral system components. Membranes are perfluorinated or partially fluorinated polymers, with side chains ending in pendant acid groups, e.g., the sulfonic acid group $-\text{SO}_3\text{H}$ (Fig. 9).

Under operation, the membrane has to contain some water, e.g., 15 water molecules per sulfonic acid group, to provide the necessary specific conductivity. This fact causes consequences for some of the system auxiliaries, as the gases, hydrogen and air, have to be humidified before flowing into the cell to sustain the hydration level in the membrane.

The acidic electrolyte requires a precious metal catalyst; hence platinum supported on carbon particles serves as electrocatalyst, with typical Pt loadings of ca. 0.1 mg/cm^2 at the anode side and ca. 0.4 mg/cm^2 at the cathode side. Platinum nanoparticles, typically a few (3–5) nanometers in diameter, are deposited on various carbon substrates (e.g., 20–40%Pt/C) by wet chemical



Fuel Cell Types and Their Electrochemistry. Figure 9
Generalized structure formula of Nafion[®] type membranes (DuPont) with $m = 1$, $l = 1$, $k = 5\text{--}7$

processes and then further processed in combination with solubilized ionomer material and binder(s) (PTFE) to yield an ink, which then is applied either to the electrolyte membrane surface (CCM, catalyst-coated membrane) or to the GDL to form the GDE.

Extensive characterization of these electrocatalysts for the fuel cell reactions has shown that the electrocatalytic activity of these nanoparticles is by a factor 10 lower than the activity of a polycrystalline platinum surface for the respective reactions, HOR or ORR [15]. This difference is not fully understood yet.

An interesting proprietary approach has been followed recently by the 3 M Company, creating a continuous electrode area covered by nanoscale Pt whiskers. Reactivity of these electrocatalytic layers is in the range of polycrystalline platinum surfaces [16].

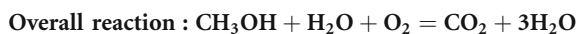
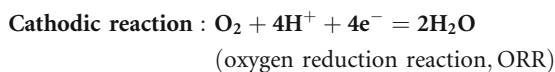
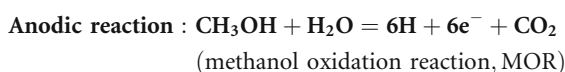
Today, pure hydrogen is considered as the ideal fuel. Due to its CO content, hydrogen derived from C-containing fuels (by steam reforming (SR), partial oxidation (POX), or autothermal reforming (ATR) followed by preferential oxidation (PROX)) would need specific measures in terms of electrocatalysis, e.g., PtRu as bifunctional electrocatalyst, to allow oxidation of CO. As an alternative, purification methods of the anode gas stream prior to entering the cell down to only a few parts per million CO (depending on operation temperature) are necessary.

The electrocatalytic layers are contacted by gas diffusion layers, allowing the gases to be passed to the interphase and the product(s), H₂O with H₂ and O₂ as reactants, which consist of carbon cloth or paper (Fig. 2). This arrangement of membrane, electrocatalytic layers, and gas diffusion layers is colloquially called membrane electrode assembly (MEA).

Most of the applications of this fuel cell type are developed for air operation, taking advantage of utilizing one reactant from the ambient environment. However, there exist also some applications in which pure oxygen is employed as oxidant. Due to the higher

partial pressure of oxygen and the absence of electrochemically inert nitrogen as the majority component in the cathodic gas stream, humidification issues of the cell can be strongly simplified [17, 18].

Methanol-Fed Fuel Cell (PEFC Type) The same concept of combination of an acidic solid polymer electrolyte and acid stable precious metal electrocatalysts can also be applied to a methanol-fed fuel cell (direct methanol fuel cell, DMFC). Methanol can be fed in liquid or vapor form, mixed with water. The methanol molecule CH₃OH is electrochemically converted at the anode:



As seen above, during MOR, one molecule of water is consumed, which is necessary to oxidize carbon in the CH₃OH molecule to CO₂, while potentially six protons are liberated. The ORR can be formulated the same way as in a hydrogen-fed cell.

The equilibrium potential for the anode reaction is 0.02 V; hence, a theoretical cell voltage close to the H₂/O₂ fuel cell should be observed. In practice, the OCV is lower, for several reasons: Due to the similarity of methanol and water as solvents (e.g., solubility parameter), methanol also penetrates into the water-swollen polymer membrane and passes to the cathode (methanol crossover), causing a mixed potential at a lower value than the oxygen potential. Further, the anode reaction requires a binary or ternary bifunctional catalyst, containing next to platinum one metal component (or two) providing the splitting reaction of the water molecule involved to liberate oxygen for the carbon oxidation at lower potentials as compared to platinum. Examples would be Ru, Sn, Mo, others, or even combinations of two of these together with platinum as the hydrogen-liberating catalyst. Intermediate oxidation species of the methanol molecule, e.g., ---COH , may adsorb and poison the platinum catalyst surface, thereby impeding the full oxidation reaction.

These kinetic losses at the anode lead to a high anode overvoltage and, therefore, the cell voltage in

the DMFC is lower at a respective value of current density as compared to a H₂-fed cell, as is the achievable power density.

Direct methanol fuel cell concepts with liquid electrolyte have also been considered in the past. One advantage of circulating liquid electrolytes is the option to cool the cell without additional cooling fluid.

High Temperature PEFC As mentioned above, H₃PO₄ can be absorbed in polybenzimidazole films to yield water-free proton-conducting membranes. The concept is strongly related to the phosphoric acid fuel cell, PAFC. These cells can be operated above 100 °C, typically in the temperature range of 160–180 °C, without humidification, allowing to simplify the balance of plant. Further, due to the higher operation temperature as compared to H₂-fed cells, CO tolerance is drastically increased. Disadvantages are the loss of H₃PO₄ during start-stop cycles, i.e., excursions to temperatures below 100 °C, where liquid product water causes leaching of the ionic component. In addition, achievable power densities are typically lower than with H₂-fed cells, due to the fact that phosphate cations strongly adsorb on platinum even at these temperatures, impeding the ORR. For more detailed information please see chapter “Benicewicz”.

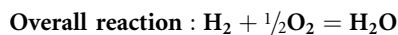
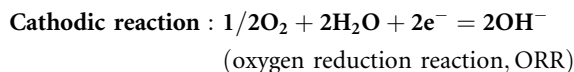
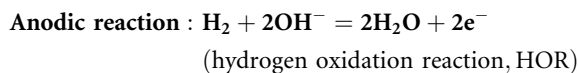
Phosphoric Acid Fuel Cell

Looking back at the development of fuel cell technologies, the PAFC technology has been the earliest one to be commercialized. Several 200 kW_e power units for co-generation have been delivered to customers over the past 20 years, mainly by the US company UTC Power [19]. Water-free phosphoric acid is absorbed in an inorganic diaphragm and contacted on both sides by platinum-containing gas diffusion electrodes. Advantages and disadvantages are very similar to the ones of the HT-PEFC.

Alkaline Fuel Cell

Alkaline electrolytes played an important role in the beginning of fuel cell development, due to its less stringent requirements for electrocatalysts. An aqueous solution of NaOH or KOH at higher concentration is absorbed in a matrix, formerly asbestos, today a composite material of an oxide and a polymer binder.

The electrochemistry of the respective HOR and ORR can be written the following way:



As described above, an alkaline cell generally has to be operated on H₂ and O₂, due to the CO₂ problem. As advantage, a non-precious metal, such as Raney nickel, can serve as electrode material. ORR in alkaline solution shows a lower overpotential loss than in acidic solution; hence, higher specific power densities (mW/cm²) can be generated in cells.

Recently, some progress in the preparation of anion exchange membranes was reported, leading to the application of the PEFC concept to alkaline fuel cells. However, performance in terms of power output and stability/lifetime is far below the acidic technology [13].

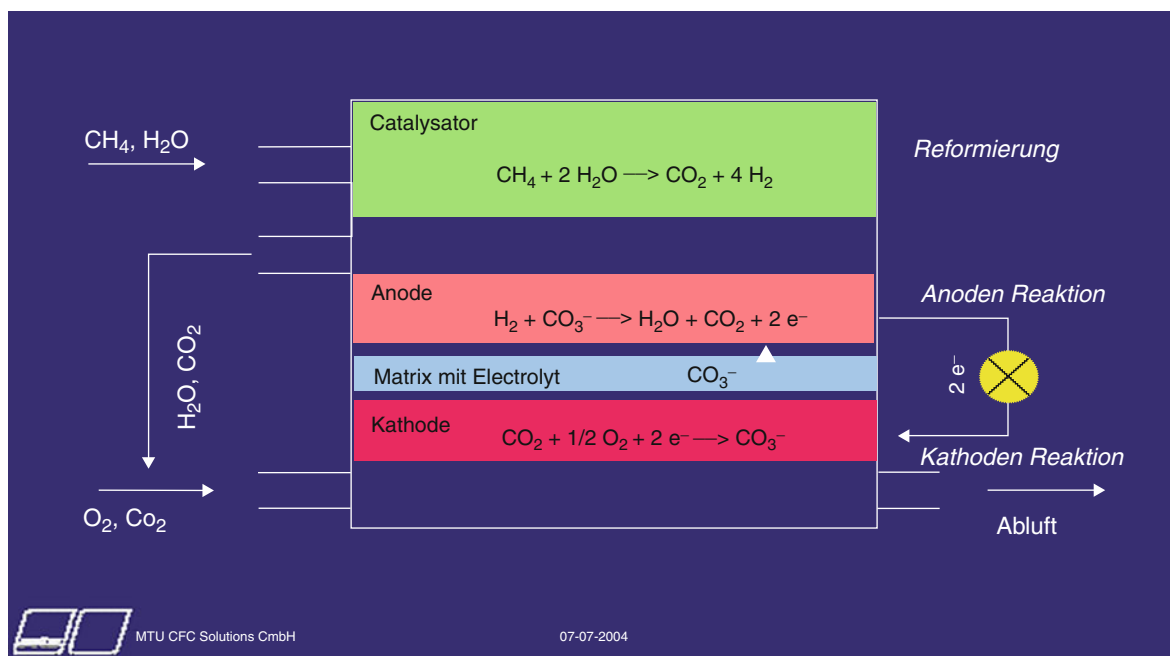
Molten Carbonate Fuel Cell

As a high temperature fuel cell, which operates in the temperature range around 650 °C, the MCFC can use the waste heat of the fuel cell reactions for internal reforming of C-containing fuels (Fig. 10). No further processing of the reformed fuel is necessary. Reformed hydrogen is oxidized at the anode, consisting of a porous nickel electrode, under consumption of the CO₃²⁻ cation, which carries the current through the electrolyte, to H₂O and CO₂. The anode reaction product CO₂ has to be recycled to the cathode compartment, where it is consumed during the ORR at a porous Ni-Oxide cathode to CO₃²⁻ and taken up by the electrolyte melt again, therefore balancing the CO₃²⁻ concentration in the Li₂CO₃/K₂CO₃ eutectic electrolyte, which is absorbed in a LiAlO₂ matrix.

MCFC systems in the power range of 250 kW have been delivered for on-site testing, using a variety of fuels, like biogas, reformed natural gas, and others [20].

Solid Oxide Fuel Cell

The electrolyte in a solid oxide fuel cell is a thin ceramic body, which is an O₂-conducting oxide material.



Fuel Cell Types and Their Electrochemistry. Figure 10

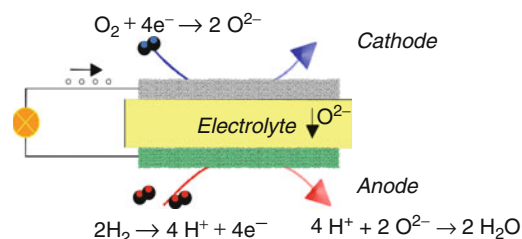
Scheme of the molten carbonate fuel cell, MCFC (Adapted from MTU CFC Solutions)

O^{2-} conduction is established by doping a stoichiometric material like ZrO_2 by Y_2O_3 , thereby generating O^{2-} vacancies in the lattice. Around 8% of Y_2O_3 are required to stabilize the cubic form of ZrO_2 at temperature of operation. A reasonable specific conductivity for this 8YSZ material is obtained in the range of and above 800°C (Fig. 11).

Electrode materials consist of a Ni/8YSZ cermet (mixture of ceramic and metal) as anode and $\text{La}_x\text{Sr}_y\text{MnO}_3$ (perovskite) as cathode.

Ohmic losses in the electrolyte can be lowered by designing cells with thin electrolyte layers, typically $10\text{ }\mu\text{m}$ thick, supported on one of the electrode structures, instead of utilizing a thicker, free-standing electrolyte layer.

For many years, also a tubular concept had been developed by Siemens Westinghouse, addressing the challenges in matching the thermal expansion coefficients of the different materials, namely for the electrolyte and the electrodes. However, the further development of this concept had been abandoned. SOFC technology development mainly addresses stationary co-generation of power and heat, at a power scale of 100 kW_e and more (e.g., Topsoe Fuel Cell,



Fuel Cell Types and Their Electrochemistry. Figure 11

Scheme of the planar solid oxide fuel cell, SOFC, and its electrochemistry (Adapted from L. Gauckler, ETH Zürich)

Denmark), as well as small scale (1 kW_e) co-generation for private homes (e.g., Hexis Switzerland).

Interestingly, low power application micro-SOFCs are developed as battery alternatives, utilizing materials which allow operation temperatures in the range of 550°C (One-Bat-Project, ETZ Zürich [21]).

Future Directions

After two decades of intensive fuel cell research and development, some of the described technologies have

come close to the market. This is particularly the case for the PEFC technology developed worldwide by automotive industry. Market introduction is planned, according to various sources, by the year 2015. However, still issues of cost and reliability remain and provide ample space for further research. Unfortunately, other technologies, e.g., the SOFC and MCFC technology, have recently seen some setbacks, due to the fact that some longstanding prominent advocates of fuel cell development have decided not to continue their effort (e.g., SOFC by Siemens, MCFC by Tognum (formerly MCFC solutions), others).

Bibliography

1. International Energy Agency (2011) Clean energy progress report, updated June 2011. International Energy Agency, Paris
2. OECD (2004) Hydrogen and fuel cells, review of National R&D Programs. International Energy Agency/OECD, Paris. ISBN 9789264108837
3. Shell International BV (2011) Shell energy scenarios to 2050, an era of volatile transitions. Shell International BV, The Hague. http://www-static.shell.com/static/aboutshell/downloads/aboutshell/signals_signposts.pdf
4. <http://www.peakoil.net/>
5. http://www.roads2hy.com/pub_download.asp?PageIndex=1
6. e.g., <http://www.sfc.com/en/>
7. Bossel U (2000) The birth of the fuel cell. European Fuel Cell Forum, Oberrohrdorf. ISBN 3-905592-06-1
8. Scherer GG (1990) Ber Bunsenges Phys Chem 94:1008
9. Hamman CH, Hamnett A, Vielstich W (1998) Electrochemistry. Wiley, Weinheim
10. Appleby AJ (1994) J Power Sources 49:15
11. Uchida H, Tanaka S, Iwahara H (1985) J Appl Electrochem 15:93
12. Haile SM (2003) Acta Mater 51:5981
13. Merle G, Wessling M, Nijmeijer K (2011) J Membr Sci 377:1
14. Kreuer K-D (2007) Proton conduction in fuel cells. In: Limbach HH, Schowen RL, Hynes JT, Klinman JP (eds) Hydrogen-transfer reactions, vol 1. Wiley, New York, pp 709–736
15. Gasteiger HA, Kocha SS, Sompali B, Wagner FT (2005) Appl Catal B Environ 56:9
16. http://www1.eere.energy.gov/hydrogenandfuelcells/pdfs/merit03/72_3m_mark_debe.pdf
17. <http://info.industry.siemens.com/data/presse/docs/m1-isfb07033403e.pdf>
18. Büchi FN, Paganelli G, Dietrich P, Laurent D, Tsukada A, Varenne P, Delfino A, Kötter R, Freunberger SA, Magne P-A, Walser D, Olsommer D (2007) Fuel Cells 07:329. doi:dx.doi.org
19. <http://www.utcpower.com/products/purecell400>
20. <http://www.fuelcellenergy.com/>
21. <http://www.nonmet.mat.ethz.ch/research/onebat>

Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels)

K. SASAKI^{1,2,3}, Y. NOJIRI¹, Y. SHIRATORI^{1,2,3}, S. TANIGUCHI¹

¹International Research Center for Hydrogen Energy, Kyushu University, Fukuoka, Japan

²International Institute for Carbon-Neutral Energy Research (WPI), Kyushu University, Fukuoka, Japan

³Faculty of Engineering, Kyushu University, Fukuoka, Japan

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Alternative Electrolyte Materials
 Alternative Electrode Materials
 Alternative Fuels
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Anode Electrode for electrochemical oxidation reactions. In solid oxide fuel cells, hydrogen-containing fuels are oxidized by oxygen ions transported through an electrolyte to form water vapor or CO₂ as the reaction products at this electrode. SOFC anodes may also act as fuel reforming catalysts when hydrocarbon-based fuels are supplied to the anodes.

Cathode Electrode for electrochemical reduction reactions. In solid oxide fuel cells, oxygen in ambient air is reduced to oxygen ions at this electrode.

Electrolyte Ionic conductor with negligible electronic conductivity, used as a membrane between an oxidation atmosphere and a reduction atmosphere. The essential material to construct electrochemical devices including fuel cells.

Fuel flexibility While low-temperature fuel cells (such as polymer electrolyte membrane fuel cells) typically use pure hydrogen gas (or hydrogen gas mixed with CO₂) as the fuel, high-temperature fuel cells

can directly use various kinds of fuels, such as carbon monoxide (CO), which is created by reforming hydrocarbons at the anode to produce hydrogen and carbon monoxide.

Impurity poisoning Even though high-temperature fuel cells have good fuel flexibility with various kinds of fuels, some minor constituents (impurities) coming from, for example, low-purity fuels, raw materials, and system components can also react with electrode materials or can be adsorbed on the electrode reaction sites, hindering electrode reactions. Such poisoning phenomena can lead to fuel cell performance degradation with time.

Mixed ionic electronic conductor Materials with both high ionic conductivity and high electronic conductivity. Such materials are often attractive for electrodes because of extension of electrochemical reaction sites beyond electrode/electrolyte interface.

Oxygen ionic conductor Materials with high oxygen ionic conductivity. Oxides with both high oxygen vacancy (or other charge carrier such as interstitial oxygen atoms) concentration and high oxygen mobility exhibit high ionic conductivities.

Protonic conductor Materials with high protonic conductivity. In the presence of oxygen vacancies, water vapor may be dissolved into oxides forming protons and exhibiting high protonic conductivity.

Definition of the Subject

While we may realize commercialization of SOFC systems in the very near future, continuous challenges will be needed to develop advanced fuel cells with higher efficiency, higher durability, better flexibility, as well as lower production cost. For realizing next-generation SOFCs, we may apply a broad range of knowledge related to (1) electrolyte and (2) electrode materials for cell development as well as (3) explore alternative fuels. This entry gives an overview on possible alternative approaches for these three important aspects to realize advanced high-performance SOFCs in a future generation.

Introduction

Fuel cells, environmentally compatible energy systems using hydrogen-containing fuels [1–5], are strongly

desired as promising key technologies to realize a sustainable low-carbon society. Compared to internal combustion, electrochemical energy conversion by fuel cells can achieve much higher efficiency. While the direct electrochemical conversion of chemical energy to electricity using fuel cells has received significant attention for several decades, mass marketing of fuel cell systems is often limited by insufficient performance and/or inadequate durability in service.

While the commercialization of fuel cell technologies has just begun in the form of, for example, stationary cogeneration power units mainly based on polymer electrolyte fuel cells for private houses, a broad variety of other application is strongly desired. Even though solid oxide fuel cells (SOFCs) are the most efficient type of fuel cells, partly due to their higher operational temperature compatible with internal reforming of hydrocarbon-based fuels, their application has often been restricted by the limited operational range of electrolyte and electrode materials [1–3]. Although many current research projects tend to focus on engineering and optimization issues related to cost reduction, durability, and reliability [6–8], in order to take SOFCs closer to full-scale commercialization, breakthroughs in materials and concepts are needed for developing viable fuel cells for broader application. It is therefore essential to consider alternative materials for electrolytes and electrodes and innovative fuel cell concepts including the use of alternative fuels for future-generation fuel cells. For example, alternative electrolyte/electrode materials for lower-temperature SOFCs are needed to improve electrical efficiency in low-power operation, tolerance against thermal cycling, as well as improve longer-term durability. They will help realize alternative fuel cells in real commercial systems. Additionally, morphological control on the nano-scale shows promise in enhancing ionic and electronic transport in nano-ionic systems [9–14], offering an additional degree of freedom in material/device design. A reversing challenge is to harness these effects to the fabrication of practical SOFCs with improved performance.

Alternative device concepts can help realize broader application of SOFCs. SOFCs for automotive applications will enable multi-fuel vehicles capable of being powered by both hydrogen and hydrocarbon fuels. The use of biogas can result in carbon-neutral energy

conversion. The use of hydrocarbon fuels and gasified coal gas fuel for SOFCs will achieve more efficient use of such fossil fuels. Portable SOFCs operating with hydrogen, hydrocarbon fuels [15, 16], and solid carbon [17] are attractive as replacements for batteries. Reversible SOFCs can act as both power units and electrolyzers, efficiently converting excess heat and electricity into hydrogen and vice versa [18–20]. If these technologies with SOFCs are realized, a considerable reduction in CO₂ emissions is possible in household, industrial, and transportation sectors.

It is therefore interesting and essential to consider such material/device issues in SOFC R&D to realize future flexible SOFCs with broader application. In this entry, an overview will be given on alternative approaches to SOFCs related to the following three critical material-related issues: electrolytes, electrodes, and fuels.

Alternative Electrolyte Materials

Electrolyte Materials: General

As the ionic conductivity of solid electrolytes is generally much lower than the electronic conductivity of electrode materials, the ionic conductivity of the solid electrolytes may determine the lowest possible operational temperature of SOFCs. Apart from high ionic conductivity, the following properties are required:

1. High ionic transference number around unity
2. Thermochemical stability in both oxidation and reduction atmospheres
3. Near theoretical density for minimum gas leakage
4. High compatibility with other materials in composite device
5. Match of thermal expansion coefficient with other components with minimal defect-induced chemical expansion

In addition, high mechanical strength and reasonable materials cost are required in the case of electrolyte-supported cells.

There are limited numbers of materials satisfying the above-mentioned criteria. Generally, Y₂O₃-stabilized ZrO₂ (YSZ) is commonly used at high temperatures, that is, $T > 800^{\circ}\text{C}$ to maintain ionic conductance, while thin-film ZrO₂-based electrolytes may

be applied for operational temperatures around 700°C . For much lower operational temperatures, alternative electrolyte materials are needed. Ionic conductivity depends strongly on the crystal structure, and oxides with high ionic conductivity have been found among open crystal structures, such as fluorite, pyrochlore, rare-earth, perovskite, and related structures [21, 22].

Oxygen Ionic Conductors

Fluorite-Type Materials

General The fluorite structure is relatively open, and it shows exceptional tolerance for high levels of atomic disorder, which may be introduced either by doping, reduction, or oxidation [23]. Among the binary oxides, CeO₂, Pr₆O₁₁, ThO₂, UO₂, and PuO₂ possess this structure in the pure state in ambient conditions. ZrO₂ and HfO₂ are stabilized in the fluorite structure by doping with divalent or trivalent oxides (Y, Sc, rare-earth oxides) [23–25]. Beyond stabilization, the addition of such dopants gives rise to the creation of oxygen vacancies responsible for the ionic conduction in these oxides.

Several fluorite-type oxides are known to be excellent oxygen ionic conductors, including δ -Bi₂O₃, CeO₂, ZrO₂, ThO₂, and HfO₂. The δ -Bi₂O₃ phase possesses an anion-deficient fluorite structure where one-fourth of the normal fluorite anion sites are vacant. This phase exhibits an order to disorder transition above $\sim 600^{\circ}\text{C}$ (which can be suppressed to some extent through doping) with a consequent increase in conductivity and has the highest oxygen ionic conductivity so far reported for a solid electrolyte [21, 22]. However, δ -Bi₂O₃ shows poor thermochemical stability, and it is easily reduced at high temperatures. ThO₂ is a radioactive material. HfO₂ has chemical similarity to ZrO₂ but is relatively expensive. Practically available fluorite-type oxides for solid electrolyte are thus limited to ZrO₂ and CeO₂. Most of the rare-earths, as well as Sc, In, and Y, are trivalent and form sesquioxides of various crystal structures with potential for solid electrolyte membranes. Most of them have the cubic C-type structure (Sm₂O₃, Dy₂O₃, Er₂O₃, Sc₂O₃, Y₂O₃, and In₂O₃) which is closely related to the fluorite structure [23], which can be derived from the fluorite structure by removing one-fourth of the oxygen atoms which lie along the four $\langle 111 \rangle$ directions [23]. SrO-doped La₂O₃ exhibits

ionic conductivity close to the Y_2O_3 -doped ZrO_2 electrolyte [21, 22]. However, most of these sesquioxides are mixed ionic and electronic conductors with a substantial electronic conductivity [23, 26–29].

ZrO₂-Based Oxide Pure ZrO_2 exhibits phase changes from monoclinic to tetragonal and cubic, with increasing temperature. Due to its volume change of ca. 5% in the phase transition from monoclinic to tetragonal, pure ZrO_2 cannot be used as a practical solid electrolyte material. Group 2 elements of the periodic table or rare-earth elements may be added to stabilize the cubic phase at high temperatures and keep this phase stable also at lower temperatures. As for stabilized ZrO_2 with dopants from the group 2 elements, CaO -stabilized ZrO_2 is widely used mainly for Nernst-type oxygen sensors [30], but it is not used for fuel cells because of its relatively low conductivity [31]. On the other hand, stabilized ZrO_2 using dopants from rare-earth elements such as Y_2O_3 , Sc_2O_3 , Yb_2O_3 , In_2O_3 , and Er_2O_3 [21, 22, 32, 33] exhibits higher conductivity suitable for SOFC [34–38]. Ionic conductivity (intragrain conductivity) of ZrO_2 doped with (a) Y_2O_3 , (b) Yb_2O_3 , (c) Er_2O_3 , (d) In_2O_3 , and (e) Sc_2O_3 measured by impedance spectroscopy is shown in Fig. 1 [33]. In addition, activation energy values for the intragrain ionic conductivity of these doped ZrO_2 , $\sigma_{\text{ion}}(\text{S/m}) \times T(\text{K}) = \sigma_0(\text{S/m}) \times \exp(-\Delta E_{\sigma}(\text{eV})/kT)$, are compiled in Table 1.

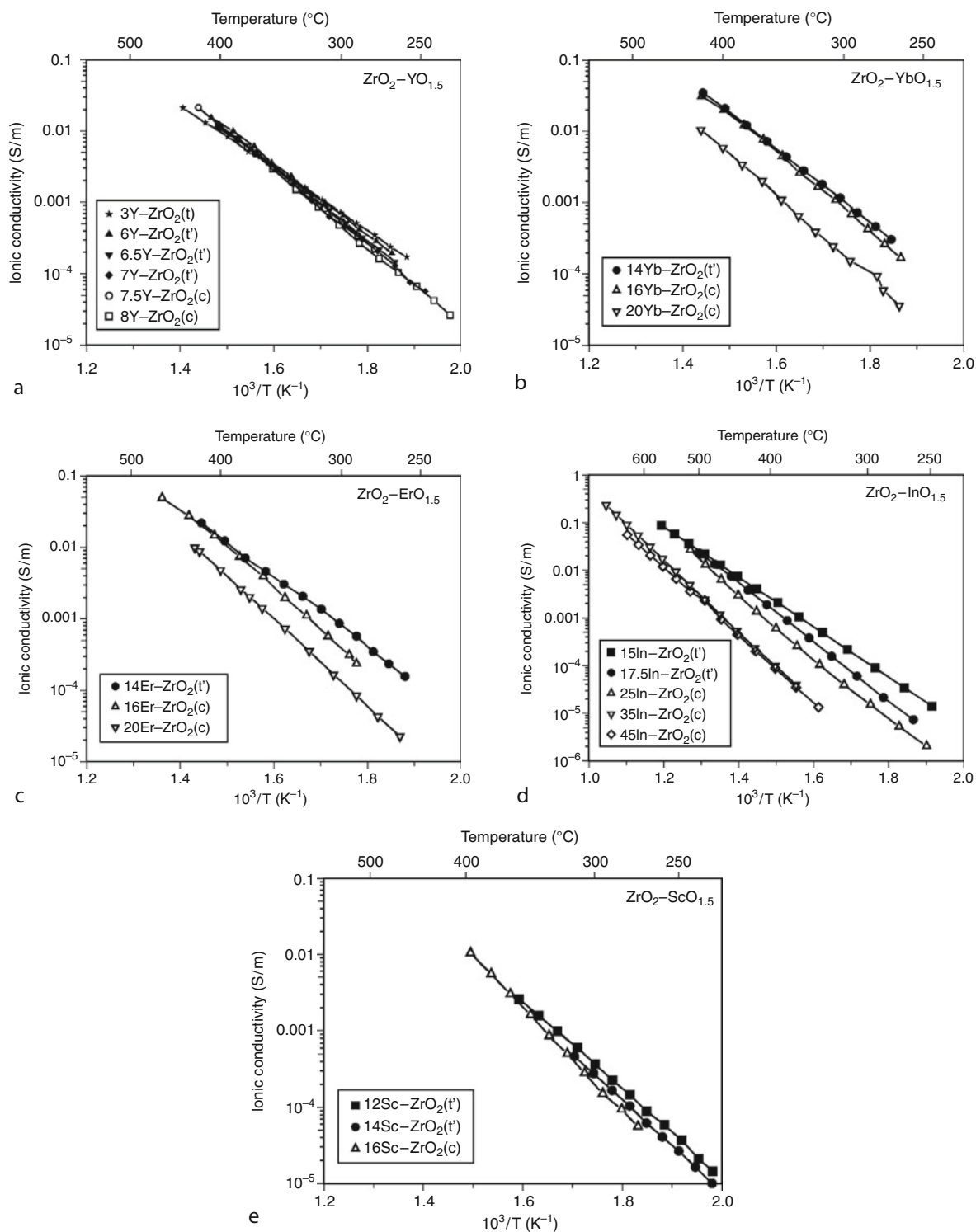
Yttria-stabilized ZrO_2 , discovered by Nernst [39], is still one of the state-of-the-art SOFC electrolyte materials which was used to demonstrate the first SOFC (and the first solid “electrolyte” fuel cell) in 1937 at ETH-Zürich [40]. Electronic defect concentrations are negligibly low [41]. As can be observed in Fig. 1, the ionic conductivity of fluorite-type oxides stabilized with hypovalent elements exhibits a maximum at a certain dopant concentration above which defect interactions occur [42–45]. As shown in Fig. 2, it is known that the peak value of the ionic conductivity becomes higher when the ionic radius of the dopant approaches that of Zr^{4+} [35], resulting in minimal distortion of the fluorite lattice. Ionic conductivity is the highest in the case of Sc^{3+} , which is used as a dopant because its ionic radius is almost identical to that of Zr^{4+} [46]. However, Sc_2O_3 is more expensive than Y_2O_3 , but Sc_2O_3 -stabilized ZrO_2 (ScSZ) has recently become

more widely used. In addition, ScSZ with 8 mol% Sc_2O_3 exhibited a high conductivity in the early stages of operation, but the conductivity decreased with operational time due to phase instability [47]. To prevent this decrease in ionic conductivity, ScSZ with 11 mol% Sc_2O_3 was preferably used, but this material had a problem of phase transition in the vicinity of 700°C [35, 48]. Currently, ScSZ with the addition of CeO_2 , such as $(\text{ZrO}_2)_{0.89}(\text{Sc}_2\text{O}_3)_{0.1}(\text{CeO}_2)_{0.01}$, is widely used to suppress this phase transition [46, 49].

In addition, tetragonal ZrO_2 and partially stabilized ZrO_2 can also be used as an electrolyte due to their better mechanical properties and their lower activation energy for ionic conductivity. Since the grain size of the tetragonal phase is generally much smaller than that of the cubic phase and grain boundaries impede oxide ion migration, the reduction of grain boundary resistivity is important [50]. Stability of the tetragonal phase for longer time periods should be considered to prevent phase transition from the tetragonal phase to the monoclinic phase when applying this material to SOFC systems.

Thin-film technology can also help to realize lower operational temperatures in SOFCs by taking advantage of short transport distance of ions. It is also possible to produce thin-film electrolytes using stabilized ZrO_2 , which enables lower-temperature operation of SOFCs because of its stability in both the oxidation and reduction atmospheres, its sinterability, and its high mechanical strength. Ultrathin YSZ electrolyte layers can be prepared on Si or glass ceramic substrates [15, 16, 51], which can realize, for example, lower temperature operation of micro-SOFCs for portable applications. Figure 3 shows an SEM cross-sectional view of a micro-SOFC prepared on a glass ceramic substrate with a YSZ electrolyte thickness of less than 1 μm , as reported by Bieberle-Hütter et al. [51].

Stabilized ZrO_2 exhibits good chemical compatibility with other materials during SOFC operation below 1,000°C. It is possible that stabilized ZrO_2 solid electrolyte reacts with other materials during the cell fabrication processes which often require high temperatures in excess of 1,300°C. YSZ does react with $\text{La}_{1-x}\text{Sr}_x\text{MnO}_{3-\delta}$ (LSM), the state-of-the-art cathode material, to produce $\text{La}_2\text{Zr}_2\text{O}_7$ or SrZrO_3 at temperatures above 1,100°C, with a consequential



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 1

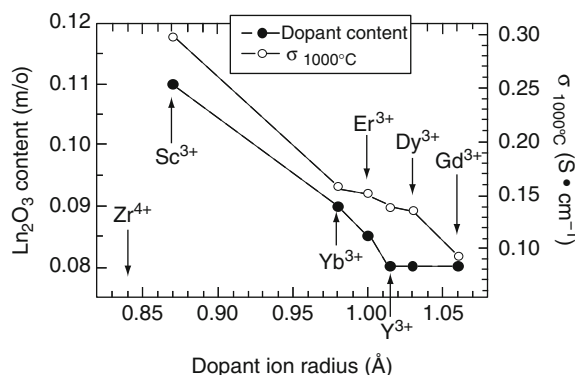
Intragrain conductivity of ZrO_2 doped with (a) Y_2O_3 , (b) Yb_2O_3 , (c) Er_2O_3 , (d) In_2O_3 , and (e) Sc_2O_3 measured by impedance spectroscopy, where t and c denote tetragonal and cubic phases [33]

Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Table 1 Activation energies and pre-exponential factors of ionic conductivity of ZrO_2 doped with trivalent metal oxides (Y_2O_3 , Yb_2O_3 , Er_2O_3 , In_2O_3 , Sc_2O_3) [33], based on the relation $\sigma_{\text{ion}}(\text{S/m}) \times T(\text{K}) = \sigma_0(\text{S/m}) \times \exp(-\Delta E_{\sigma}(\text{eV})/kT)$, where t and c denote tetragonal and cubic phases

Dopant	mol%(phase)	Intragrain conductivity	
		$\Delta E_{\sigma}(\text{eV})$	σ_0
$\text{YO}_{1.5}$	5.8 (t) ^{fc}	0.91	5.11×10^7
	11.2 (t')	1.04	5.49×10^8
	12.0 (t')	1.06	6.93×10^8
	12.7 (t')	1.10	1.34×10^9
	13.5 (c)	1.09	1.20×10^9
	14.3 (c)	1.12	1.69×10^9
$\text{Yb}_{1.5}$	14(t')	1.07	1.43×10^9
	16 (c)	1.13	3.66×10^9
	20 (c)	1.19	3.74×10^9
$\text{ErO}_{1.5}$	14 (t')	1.02	3.86×10^8
	16 (c)	1.20	7.51×10^9
	20 (c)	1.25	7.74×10^9
$\text{InO}_{1.5}$	15 (t)	0.80	6.39×10^6
	15 (t')	1.10	2.99×10^8
	17.5 (t')	1.28	4.29×10^8
	25 (c)	1.35	7.51×10^9
	25 (c) ^{fc}	1.22	5.26×10^9
	35 (c)	1.56	3.47×10^{10}
	35 (c) ^{fc}	1.14	1.57×10^8
	45 (c)	1.48	8.22×10^9
	45 (c) ^{fc}	1.26	1.83×10^9
$\text{ScO}_{1.5}$	12 (t')	1.20	6.94×10^9
	14 (t')	1.24	1.20×10^{10}
	16 (c)	1.40	2.44×10^{11}

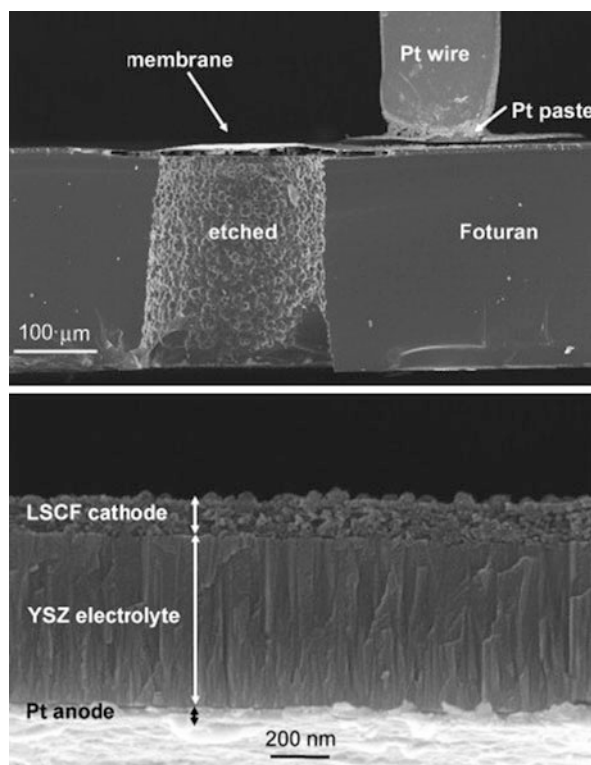
fc furnace cooling at a rate of $10^\circ\text{C}/\text{min}$. Others: air quenching

increase in internal cell resistance [52–54]. In the case of LaCoO_3 - or LaFeO_3 -based oxide cathode materials, reaction with stabilized ZrO_2 becomes worse, so that CeO_2 -based oxides such as Gd_2O_3 -doped CeO_2 are generally used as a buffer layer between the ZrO_2 -based



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 2

Relation between dopant ionic radius, conductivity, and content of dopant, reported by Arachi et al. [35] (Reproduced by permission of Elsevier)

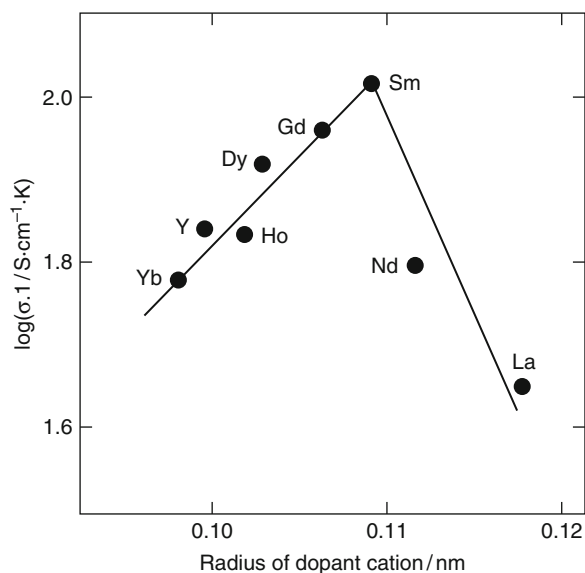


Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 3

SEM cross-sectional view of a micro-SOFC prepared on a glass ceramic substrate, as reported by Bieberle-Hütter et al. [51] (Reproduced by permission of Elsevier)

electrolyte and such cathode materials to prevent interfacial reactions from forming [3, 55].

CeO₂-Based Oxide CeO₂-based oxide, partially substituted by a hypovalent rare-earth oxide, exhibits high ionic conductivity in a wider range of compositions in comparison to stabilized ZrO₂. This is because even pure CeO₂ has the fluorite-type cubic structure rendering stabilization unnecessary. Like ZrO₂, ionic conductivity of CeO₂ also varies with the substitution of elements. It shows high ionic conductivity when it is partially substituted with alkaline-earth elements such as Ca and Sr [56, 57], and when partially substituted with a rare-earth element, even higher conductivity. As shown in Fig. 4, there is a systematic trend of the conductivity with ionic radius of the substituting element giving the highest conductivity with Sm [58]. Pure Sm₂O₃ has a rare-earth C-type structure, and the similarity of structures between Sm₂O₃ and CeO₂ may lead to high stability. High ionic conductivity can be obtained in a wide range of substitution ratios with Sm₂O₃.



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 4

Relationship between the dopant ion radius and the ionic conductivity of doped CeO₂ at 800°C, as reported by Eguchi et al. [58] (Reproduced with permission from Elsevier)

Sm₂O₃-doped CeO₂, usually abbreviated as SDC, with 20 mol% partial substitution of Sm₂O₃, exhibits the highest ionic conductivity. In addition, CeO₂ partially substituted with Gd₂O₃ is also widely used as an electrolyte and, as previously mentioned, interlayer between the stabilized ZrO₂ electrolyte and the cathode, although it shows slightly lower conductivity than SDC [59]. More recently, double doping with Sm and Nd exhibits one of the highest ionic conductivities reported for ceria [60].

When doped CeO₂ is applied as a solid electrolyte for SOFCs, a change in valence state of Ce⁴⁺ to Ce³⁺ in a reducing atmosphere at the anode results in two main drawbacks. First, it causes n-type electronic conduction in the solid electrolyte which leads to a cell voltage drop. Precisely speaking, when CeO₂ is reduced, both oxygen vacancy and electron are introduced simultaneously. Since the mobility of electrons is much higher than that of oxygen ions, the transference number, that is, the fraction of ionic conductivity from the total, decreases from unity. Second, when lattice oxygen is removed by reduction, the lattice expands, known as chemical expansion. A steep oxygen potential gradient appears in the CeO₂ electrolyte from reducing anode to oxidizing cathode under the operating conditions of SOFCs, and a large mechanical stress in the electrolyte can occur resulting in fracture of the cell [61].

These issues can be overcome partially by using stabilized ZrO₂ as an interlayer between the anode and the electrolyte [62–65]. However, since there is a thermal expansion mismatch between the ZrO₂-based solid electrolyte and the CeO₂-based interlayer, this often causes microscopic fracture [64, 65]. GDC with an addition of Pr (Ce_{0.8}Gd_{0.18}Pr_{0.02}O_{2-δ}) showed enhanced reduction resistance but with the same level of ionic conductivity as SDC [66]. In addition, there have been some attempts to realize lower-temperature operation by fabricating ultrathin membranes of CeO₂ electrolyte materials which contribute to reduce electrical resistance. Electronic conductivity, caused by reduction, decreases with decreasing temperature, but ionic conductivity also decreases simultaneously. Although making thinner solid electrolytes is accompanied with certain difficulties, CeO₂-based electrolytes can be operated at lower temperatures below 600°C, where the influence of electronic conduction caused by the reduction may not be a fatal problem [67].

For example, Oishi et al. [67, 68] have reported stable operation of metal-supported CeO₂-based SOFCs, composed of a porous ferritic stainless steel support, a Ni-Ce_{0.8}Y_{0.2}O_{1.9} anode, a thin Ce_{0.9}Gd_{0.1}O_{1.95} electrolyte, and a La_{0.6}Sr_{0.4}Co_{0.2}Fe_{0.8}O₃ cathode, over 1,200 h at 600°C. CeO₂-based electrolyte with a thickness of about 10 μm allows operation of SOFCs at rather low temperatures and selection of ferritic steel as a support material, resulting in enhancement of mechanical strength of cells, improvement of redox stability of cell stacks, and minimization of the use of expensive electrode and electrolyte materials [67, 68]. Considering transport in thin-film CeO₂-based electrolytes, nano-size effects [12] and mixed conduction [62] should be taken into account to tailor transport phenomena in SOFCs.

Pyrochlores and Other Fluorite-Type Oxides (Y, Nb, Zr) O_{2-δ} Pyrochlore-type oxides have the general formula A₂³⁺B₂⁴⁺O₇ in which A is a rare-earth element such as Gd or Y, and B is Ti or Zr. Gd₂Zr₂O₇ is the typical composition of pyrochlore-type oxides and can be considered as fluorite-type in which ionic defects are regularly arranged. The defect structure and the mixed conductivity can be controlled by the value of x in, for example, Gd₂(Zr_xTi_{1-x})₂O₇, which is abbreviated as GZT [69]. When the ionic radius of rare-earth elements for the A site is larger than that of Gd, the structure changes from highly defective fluorite to pyrochlore [70, 71].

In the Gd₂Ti₂O₇ series, relatively high ionic conductivity is obtained at the 20 mol% substitution of Gd sites with Ca [72], though its use as an SOFC solid electrolyte is precluded by its electronic conductivity in air, with an ionic transference number lower than that of YSZ [73]. The thermal expansion coefficient of Gd₂Ti₂O₇ (10.4–10.8 × 10⁶ K⁻¹) [73, 74] is close to that of LSGM (10.4 × 10⁶ K⁻¹) [75] and 10ScSZ (10.9 × 10⁶ K⁻¹) [76].

In addition, the solid solution of Y₄NbO_{8.5} and (Y, Nb, Zr)O_{2-δ} also has the fluorite-related structure, and these materials have stimulated research interest due to their oxide ion conduction [77, 78]. Electrical conductivity of Y₄NbO_{8.5} is independent of the oxygen partial pressure, and it is recognized as an ionic conductor [78]. The ionic conductivity is low, enhancement of conductivity by the addition of Zr

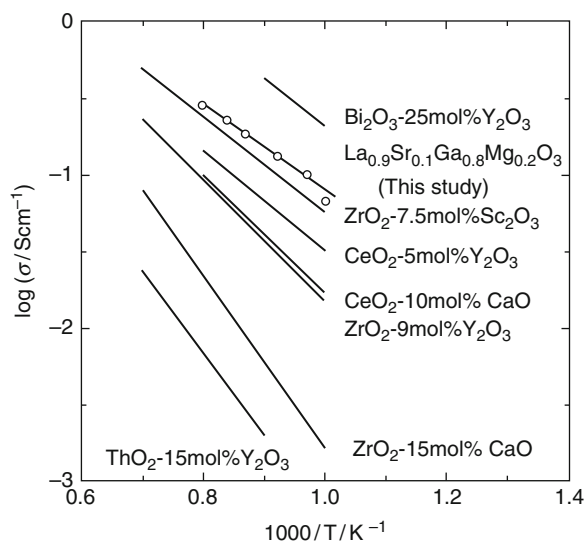
was reported, and further improvement of ionic conductivity may be expected [77].

Perovskite-Related Oxides

General Perovskite-type oxides have the general formula ABO₃. The total valence of the A metal and the B metal ions is 6, and various combinations are possible. However, there is a limit to the ionic radius (r) of the constituent metal ions (A and B) and the oxygen ion (O), as the tolerance factor *t* ($t = \frac{R_A + R_O}{\sqrt{2} \cdot (R_B + R_O)}$) should satisfy 0.75 ≤ *t* ≤ 1 [79, 80]. Oxides related to LaGaO₃, LaAlO₃, and Ba₂In₂O₅ have been well examined as solid electrolyte materials for SOFCs. Takahashi and Iwahara [81] already showed in 1970s that selected perovskite solid solutions exhibited high ionic conductivity around 5 × 10⁻³ S/m at 800°C [21, 22].

LaGaO₃-Based Oxides Oxygen ionic conductivity in perovskite-type-oxide structures has been extensively investigated by Takahashi and Iwahara in their pioneering and detailed studies [81]. High ionic conductivity was observed in the Ti series of perovskite-type oxides. Despite detailed studies, perovskite-type titanate showed poor ionic conductivity compared to that of YSZ. Furthermore, hole conduction was relatively high, so that the ionic transference number was less than unity in oxidizing atmospheres. They also investigated perovskite-type oxides in which trivalent cations were added to both the A site and the B site. LaCoO₃ and LaFeO₃ are known as such kinds of oxides, but these oxides are mixed conductors with predominately high hole conductivity. However, LaAlO₃, with trivalent metal ions on the B site, showed relatively high ionic conductivity [81].

Building on these results, Ishihara et al. [82] attempted to increase lattice cubic volume by applying Ga to the B site which has a slightly larger ionic radius than Al in the La-based perovskite oxides (Fig. 5). They added various dopants, and the highest oxygen ionic conductivity was obtained in the case where 20 mol% of Sr was added to the A site, which has an approximately equal ionic radius to La [82]. As for the dopant on the B site, the highest ionic conductivity was obtained when 20 mol% of Mg was added to the B site (La_{0.8}Sr_{0.2}Ga_{0.8}Mg_{0.2}O_{2.6} [83]), since Mg



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 5

Comparison of the oxygen ionic conductivity of $\text{La}_{0.9}\text{Sr}_{0.1}\text{Ga}_{0.8}\text{Mg}_{0.2}\text{O}_{3}$ with conventional oxide ionic conductors with fluorite structures, compiled by Ishihara et al. [82] (Reproduced with permission from American Chemical Society)

is the only element which satisfies the criteria of both ionic radius and low valence number. $\text{La}_{0.8}\text{Sr}_{0.2}\text{Ga}_{0.8}\text{Mg}_{0.2}\text{O}_{2.6}$ (LSGM) shows the same level of ionic conductivity as GDC at temperatures above 650°C . Electronic conductivity does not become predominant within a wide range of atmospheres from pure oxygen to hydrogen-containing gases [84, 85].

While LSGM exhibits superior properties as an SOFC solid electrolyte, a phase transition from pseudo cubic (monoclinic) to tetragonal was observed at temperatures around 700°C . In order to give high ionic conductivity to this material even below 650°C , the B site was partially substituted with Co, which has a similar ionic radius to Mg ($\text{La}_{0.8}\text{Sr}_{0.2}\text{Ga}_{0.8}\text{Mg}_{0.115}\text{Co}_{0.085}\text{O}_{3-\delta}$, denoted as LSGMC) [84]. The crystalline phase at high temperature was then stabilized, and the phase transition did not occur at temperatures around 700°C , resulting in higher ionic conductivity than that of GDC within a wide temperature range from 500°C to $1,000^{\circ}\text{C}$ [85]. It was revealed that oxygen ions tend to be trapped around Mg in

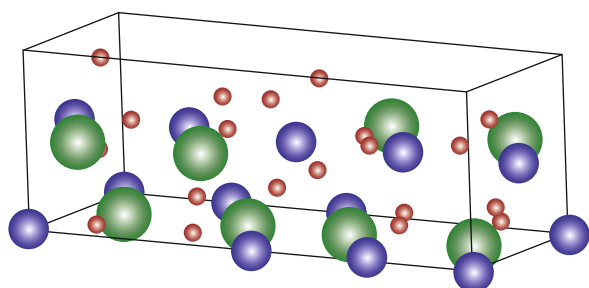
LSGM [86], and the addition of Co on the B site is effective for preventing such trapping at lower temperatures.

When an LaGaO_3 -based oxide is applied as a solid electrolyte for SOFCs, reactivity and/or interdiffusion with electrode materials becomes a critical issue. When LSM was used as the cathode material, interdiffusion of Mn into LSGM was observed [87]. When LSCF was used as the cathode material, interdiffusion of Co and Fe into LSGM was also seen [88–90]. A solid-state reaction also occurs at the interface between the Ni-based anode and the LSGM electrolyte due to the similarity of Ni to Co and Mn in the electrolyte materials [91]. There have been some attempts to prevent such solid-state reactions, for example, by optimizing deposition procedures and by applying a CeO_2 interlayer between the LSC cathode and the LSGM electrolyte.

Ba₂In₂O₅-Based Oxides The brownmillerite-type oxide has the general formula $\text{A}_2\text{B}_2\text{O}_5$, with the crystal structure presented in Fig. 6. Alkaline-earth metal ions such as Ba occupy the A site, and trivalent typical metal ions such as In occupy the B site [92–99]. At temperatures below 1,140 K, it behaves as a mixed conductor which shows both p-type electronic conductivity and oxygen ionic conductivity due to its oxygen excess [92]. When it is heated to 1,140–1,230 K, ionic conductivity increases, which is associated with the phase transition to the disordered perovskite structure, as shown in Fig. 7 [93]. The ionic transference number also rises drastically to unity.

For example, in the case of $\text{Ba}_2\text{In}_2\text{O}_5$, the typical brownmillerite-type oxide, the disordered perovskite structure can be stabilized, and ionic conductivity in an intermediate-temperature range can be improved by substituting In with another highly charged cation such as Zr, Hf, and Sn [94, 95]. A similar effect can be obtained by the substitution of Ba, as shown in Fig. 8 [93, 96].

$\text{Ba}_2\text{In}_2\text{O}_5$ is stable in the absence of CO_2 but decomposes in the presence of CO_2 in a reducing atmosphere ($p(\text{O}_2) = 10^{-9}$ – 10^{-13} atm), accompanied by the formation of BaCO_3 . However, in a strongly reducing atmosphere of much lower oxygen partial pressure, In^{3+} is reduced to In^+ , and the brownmillerite structure cannot be maintained anymore [92, 95, 97–99].



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 6

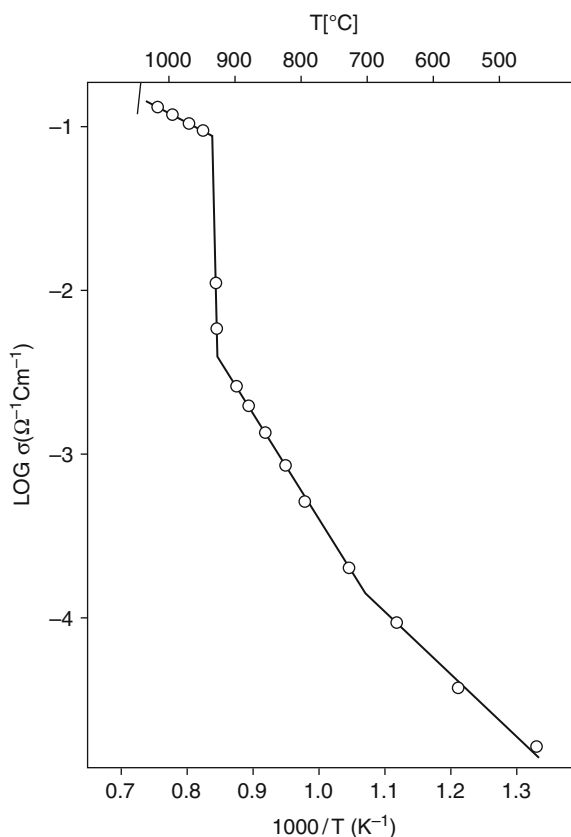
Crystal structure of $\text{Ba}_2\text{In}_2\text{O}_5$ (Green: Ba, Blue: In, Red: O)

Apatite-Type Oxides

General Apatite is a generic name for minerals with the composition $\text{A}_{10}(\text{MO}_4)_6\text{X}_2$ [100–102]. Generally, the A site can be occupied by alkaline, alkaline-earth, or rare-earth metal ions. The M site can be occupied by elements such as P, Si, and Ge which can have a coordination number of 4. The M site can be partially substituted with transition metal ions. The X site can be occupied by O, OH, F, Cl, and CO_3 , etc. Moreover, the A-site cations are divided into two kinds: one occupies the 4f site where the coordination number is 7, and the other occupies the 6h site where the coordination number is 9. Hydroxide apatite $\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})$ is a typical example. It is the principal ingredient of human bone and is practically used for artificial bone [101–103].

$\text{La}_{9.33+x}\text{Si}_6\text{O}_{26+3x/2}$ ($x = 0\text{--}0.67$) $\text{La}_{9.33}\text{Si}_6\text{O}_{26}$ and $\text{La}_{10}\text{Si}_6\text{O}_{27}$ have often been studied for their crystallographic interest. It is already known that lattice volume will continuously decrease with the ionic radius of lanthanide ions from La to Lu [102, 104]. In the $\text{Re}_{9.33+x}\text{Si}_6\text{O}_{26}$ system, oxides with $\text{Re} = \text{La}\text{--}\text{Dy}$ belong to the hexagonal system apatite structure ($\text{P6}_3/\text{m}$) and oxides with $\text{Re} = \text{Y}\text{--}\text{Yb}$ to the monoclinic system apatite structure (I2/a). Specifically, $\text{La}_{10}\text{Si}_6\text{O}_{27}$ is not regarded as the apatite phase. In the case of silicate and germanate, the oxides with the amount of La from 8 to 9.33 exhibit the apatite phase [105].

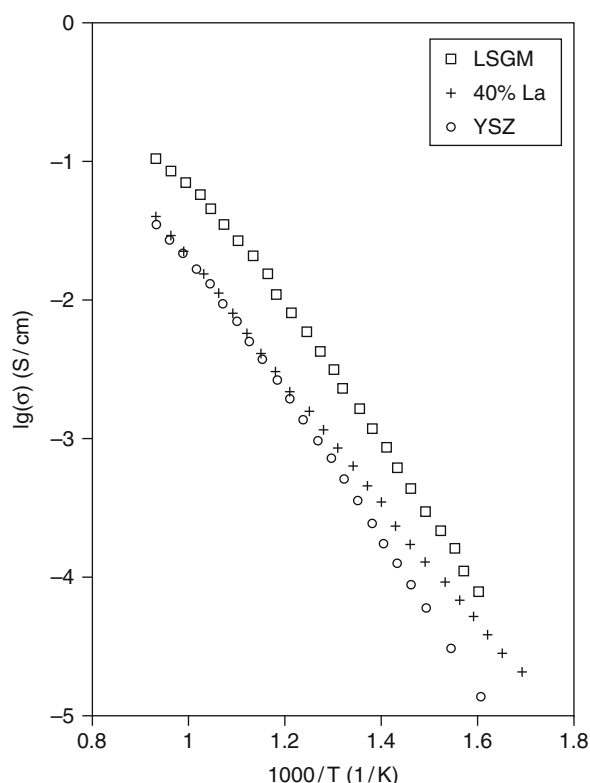
The crystal structure of this system is shown in Fig. 9. The La 6h site is fully occupied, while the La 4f site is not fully filled. $\text{La}_{9.33}\text{Si}_6\text{O}_{26}$ does not have



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 7

Ionic conductivity of $\text{Ba}_2\text{In}_2\text{O}_5$, reported by Goodenough [93] (Reproduced with permission from Elsevier)

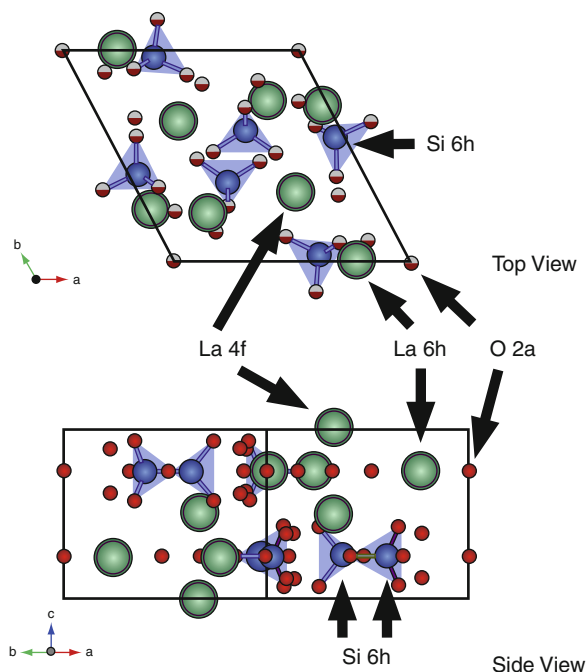
interstitial oxygen ions, but the conductivity improves by partially substituting the La 4f site or by preparing a slightly La-deficient sample ($\text{La} = 9.0$) to introduce oxygen vacancies into the 2a site [105, 106]. For $x > 0$, oxygen ions are introduced into the interstitial site, and 1 mol of oxygen is introduced into the unit cell at $\text{La} = 10$ ($x = 0.67$). When the number of interstitial oxygen ions increases, the conductivity also increases [107]. The ionic conductivity in this system is strongly anisotropic, as its crystallographic structure is. A single-crystal conductivity measurement of $\text{Nd}_{9.33}\text{Si}_6\text{O}_{26}$ has revealed that the conductivity along the c-axis is more than 10 times higher than that in the a–b plane [108]. The anisotropy of this conductivity disappears at around 800°C [109].



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 8

Comparison of the oxygen ionic conductivity of $\text{La}_{0.4}\text{Ba}_{0.6}\text{InO}_{2.7}$ with $\text{La}_{0.9}\text{Sr}_{0.1}\text{Ga}_{0.8}\text{Mg}_{0.2}\text{O}_{2.85}$ and YSZ, reported by Goodenough [93] (Reproduced with permission from Elsevier)

As solid electrolytes, the La/Si oxides with $\text{La} > 9.33$ contain interstitial oxygen ions, which can conduct in the channel (2a site) along the c-axis. The conduction mechanism is quite different from that of other oxide electrolytes in which oxygen vacancies are the mobile charge carriers. For example, in $\text{La}_{10}\text{Si}_6\text{O}_{27}$, the activation energy of ionic conductivity via oxygen interstitials is lower than that of YSZ, and the ionic conductivity of this La/Si oxide is higher than that of YSZ at and below 923 K, as described in Fig. 10 [110]. Moreover, it is known that the transference number is unity, and that this material remains stable in humidified hydrogen and in CO_2 -containing atmospheres at and below 1,073 K, while p-type conduction can appear at and above 1,173 K [111, 112]. The ionic conductivity decreases, and its activation energy increases with decreasing ionic radius of the rare-earth metal ions in



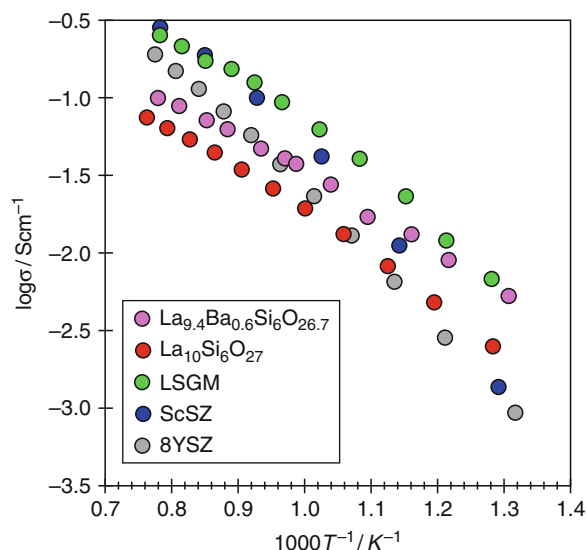
Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 9

Crystal structure of $\text{La}_{10}\text{Si}_6\text{O}_{27}$ [110] (Green: La, Blue: Si, Red: O) (Reproduced with permission from Elsevier)

the La site [113, 114]. The substitution sites of dopants should be taken into account in considering ionic conductivities of such materials [106, 107, 110, 115, 116].

These materials are based on inexpensive La_2O_3 and SiO_2 and can be synthesized at relatively low temperatures, but cannot exhibit high conductivity and density without sintering at and above 1,873 K [117]. The thermal expansion coefficient of $\text{La}_{10}\text{Si}_6\text{O}_{27}$ is $9.7\text{--}10.3 \times 10^{-6} \text{ K}^{-1}$. This value is very close to that of YSZ ($10.5 \times 10^{-6} \text{ K}^{-1}$) and GDC ($11.8 \times 10^{-6} \text{ K}^{-1}$) [118, 119], indicating that similar oxide electrodes may be used. There are a few reports on fuel cell electrochemical performance using a $\text{La}_{0.9}\text{Sr}_{0.1}\text{CoO}_{3-\delta}$ cathode with a Ni-SDC cermet anode [110, 116, 119]. The results show 50% better performance than that with Pt electrodes [119]. The interdiffusion of La and Sr in the electrode and in the electrolyte can be technological issues, as observed also in the LSGM-electrolyte-based cells.

$\text{La}_{9.33+x}\text{Ge}_6\text{O}_{26+3x/2}$ ($x = 0\text{--}2.67$) It has been already reported that La/Ge oxides exhibit a phase transition



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 10

Comparison of the oxygen ionic conductivity of lanthanum silicate with $\text{La}_{0.9}\text{Sr}_{0.1}\text{Ga}_{0.9}\text{Mg}_{0.1}\text{O}_3$ and oxides with the fluorite structure [110]

from triclinic to hexagonal phase at around 1,000 K [120]. This phase transition disappears by substitution of La by Sr [121]. A similar phase transition is observed in La_2GeO_5 [122]. The ionic conductivity of $\text{La}_x\text{Ge}_6\text{O}_{12+1.5x}$ exhibits the highest value in the apatite region, while this result is not the same as in the lanthanum silicates [105, 119, 123]. Due to the volatility of Ge, La_2GeO_5 is difficult to synthesize in the stoichiometric ratio of La/Ge [124].

Protonic Conductors

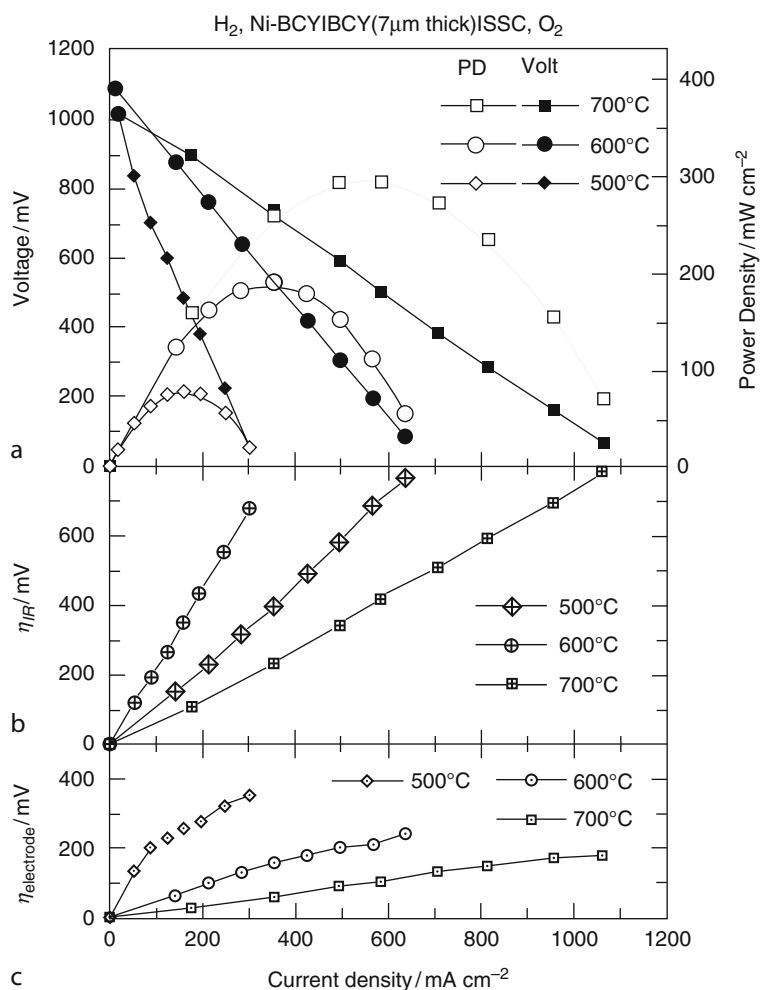
General In general, most ionic conductors currently used for solid electrolytes in SOFCs are oxygen ionic conductors. However, the use of proton-conducting oxides allowing for lower-temperature operation is anticipated to have a beneficial effect on durability at long-term operation. There are at least two approaches for this aim from the electrolyte materials selection viewpoint. One approach is to increase the oxygen ionic conductivity of the electrolytes, and another approach is to use a proton-conducting oxide which exhibits smaller temperature dependence (i.e., lower activation energy) of the ionic conductivity.

There are many kinds of proton-conducting oxides widely known in the literature, for example, by Kreuer [125]. In a very early study, C. Wagner had already reported the presence of protons in stabilized ZrO_2 [126]. Hydrated oxides such as $\text{WO}_3 \cdot n\text{H}_2\text{O}$, phosphates such as $\text{Zr}(\text{HPO}_4)_2$, β -alumina, β'' -alumina, zeolites, and heteropoly acid are good examples of proton-conducting oxides [9, 125, 127–130]. Acid salts such as CsHSO_4 are also considered as electrolyte materials for fuel cells operating at lower temperatures [131, 132].

Various perovskite-type oxides [133, 134], as well as $\text{Ba}_2\text{In}_2\text{O}_5$ [135–137], and $\text{La}_{10}\text{Ge}_6\text{O}_{27}$ [120, 138, 139] exhibit protonic conductivity. Among these materials, the perovskite-type oxides show outstanding protonic conductivity. To mention a few, SOFCs with $\text{BaCe}_{0.8}\text{Y}_{0.2}\text{O}_{3-\delta}$ solid electrolytes can be operated at low temperatures above 400°C [140, 141], and a power density of 100 mW/cm² at 500°C has been achieved (Fig. 11, [141]).

Perovskite-Type Oxides Regarding protonic conduction in perovskite-type oxides (ABO_3), Takahashi and Iwahara [81] discovered the interesting phenomenon that protonic conduction appeared in SrCeO_3 at high temperatures, which was associated with the presence of oxygen vacancies [125]. They revealed that BaCeO_3 has high hydrogen solubility by using deuterium absorption measurements [142]. This material exhibited mixed conductivity of oxygen ions and protons, which was strongly affected by the Ba/Ce ratio [143, 144]. When the Ba/Ce ratio was more than 1.05, such oxides exhibited high protonic conductivity. When the Ba/Ce ratio was less than 1, it behaved as an oxygen ionic conductor, although the conductivity was relatively low [145]. They also investigated the effect of partial substitution in detail [146–153].

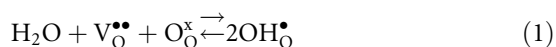
The most important reaction leading to the formation of protonic defects at moderate temperatures is the dissociative absorption of water which requires the presence of oxide ion vacancies $\text{V}_{\text{O}}^{\bullet\bullet}$. Those defects may be formed intrinsically by varying the ratio of main constituents (e.g., in $\text{Ba}_3\text{Ca}_{1+x}\text{Nb}_{2-x}\text{O}_{3-\delta}$ [154] or $\text{Ba}_2\text{YSnO}_{5.5}$ [155]), or they may be formed extrinsically to compensate for an acceptor dopant R. In the case of perovskite-type oxides, one generally employs a substitution of up to 25% of the B-site cation by



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 11

(a) I - V characteristics (closed symbols) and power density curves (open symbols), (b) ohmic loss and (c) electrode overpotentials (anode + cathode) of a H_2 - O_2 proton conductor thin film SOFC operated at 700–500°C, reported by Matsumoto et al. [141] (Reproduced by permission of Elsevier)

a lower-valence cation (i.e., $\text{A}(\text{B}_{1-x}\text{R}_x)\text{O}_{3-\delta}$ in the following denoted R:ABO_3). In order to form protonic defects, water from the gas phase dissociates into a hydroxide ion filling an oxide ion vacancy and a proton forming a covalent bond with a lattice oxygen. In Kröger–Vink notation, this reaction is written:



by which two hydroxide ions substituting for oxide ions, that is, two positively charged protonic defects ($\text{OH}_{\text{O}}^{\bullet}$), are formed. The hydroxyl ions form dynamical

hydrogen bonds with neighboring lattice oxygen, and proton transfer through these hydrogen bonds combined with the rotational diffusion of this defect allows the protonic charge to migrate [156].

For perovskite-type oxides, the proton solubility decreases, but the stability to CO_2 increases as the basicity of B-site ions decreases in the order of $\text{Ce} \rightarrow \text{Zr} \rightarrow \text{Sn} \rightarrow \text{Nb} \rightarrow \text{Ti}$ [156]. The best compromise for high proton conductivity and chemical stability is found for Y-doped BaZrO_3 , but unfortunately, this material exhibits large grain boundary resistance, and it still reacts slightly with CO_2 under certain

conditions [157]. Chemical stability against CO_2 has been demonstrated recently with $\text{BaCe}_{0.3}\text{Zr}_{0.5}\text{Y}_{0.2}\text{O}_{3-\delta}$ by Fabbri et al. [158].

Proton concentration is proportional to the solubility of water into oxygen vacancies. Such a protonic conductor tends to lose water at high temperatures, which results in a decrease in the carrier density and also a decrease in conductivity. Figure 12 shows the amount of water dissolution measured by the gravimetric analysis for protonic conductors [156]. The perovskite-type proton-conductive oxides have enough carrier density in the temperature range from ca. 300 to 500°C. On the other hand, mobility changes with temperature according to the Arrhenius equation with activation energy approximately from 0.3 eV to 0.6 eV. As a result, these protonic conductors exhibit their maximum conductivity at certain temperatures as shown in Fig. 13 [156].

It is extremely important that the ionic transference number is high enough to be used as a solid electrolyte in fuel cells. The Ce^{4+} cation has a tendency to be reduced easily, so that electronic conduction caused by the reduction of Ce^{4+} ions becomes a problem in the case of the CeO_2 -containing electrolyte. BaCeO_3 -based oxides contain Ce ions, exhibiting certain

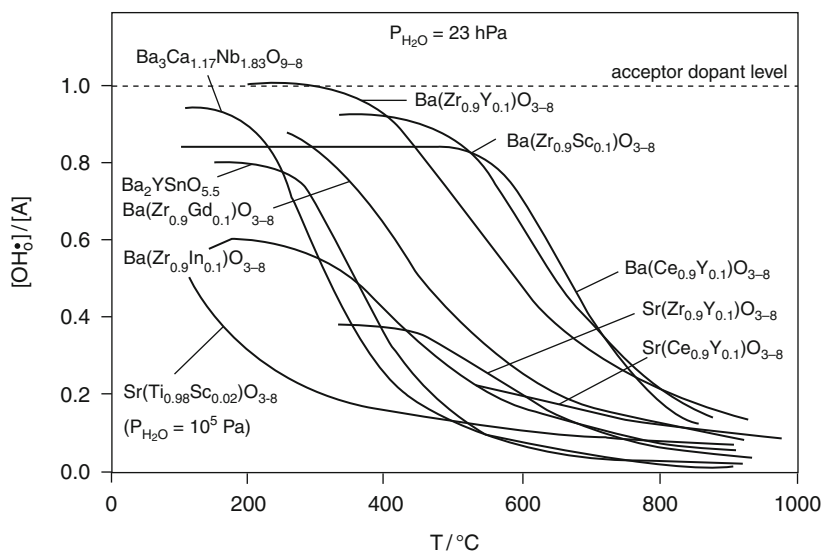
electronic conductivity by reduction [157]. Transport parameters including activation enthalpies of hole and proton conduction have been reported [157, 159].

It is commonly believed that this is because the band gaps of the perovskite-type protonic conductors are considerably large, ranging from 4.5 to 5 eV or even larger [160]. On the other hand, hole conduction appears in an oxidizing atmosphere. However, the temperature dependence of the hole conductivity is larger than that of the protonic conductivity. Oxygen vacancies are filled with water dissolved in various oxides, as shown in Fig. 12, at temperatures below 500°C, leading to a high protonic conductivity in an oxidizing atmosphere.

Alternative Electrode Materials

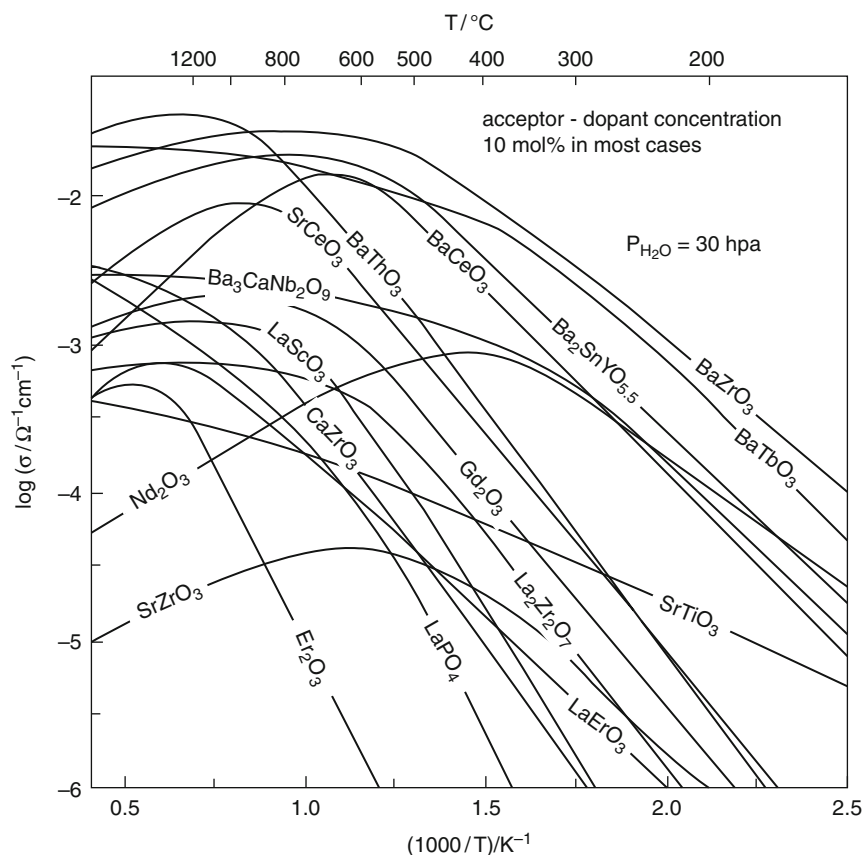
Cathode Materials

Current Cathode Materials The majority of investigations related to SOFC cathode materials have been concentrated on oxides with the perovskite structure [2–4, 161]. The oxides commonly used are SrO-doped LaMnO_3 with an A-site deficiency ($(\text{La}_{1-x}\text{Sr}_x)_{1-y}\text{MnO}_{3\pm\delta}$; $x = 0.1\text{--}0.5$, $0 \leq y < 0.1$). For example, $\text{La}_{0.86}\text{Sr}_{0.14}\text{MnO}_3$ exhibits an electronic



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 12

Normalized hydrogen isobars ($p_{\text{H}_2\text{O}} = 23$ hPa) for different perovskites, reported by Kreuer [156]. For BaZrO_3 -based compositions, data for different hypovalent dopants (Y, Sc, Gd, In) are included (Reproduced by permission of Annual Reviews)

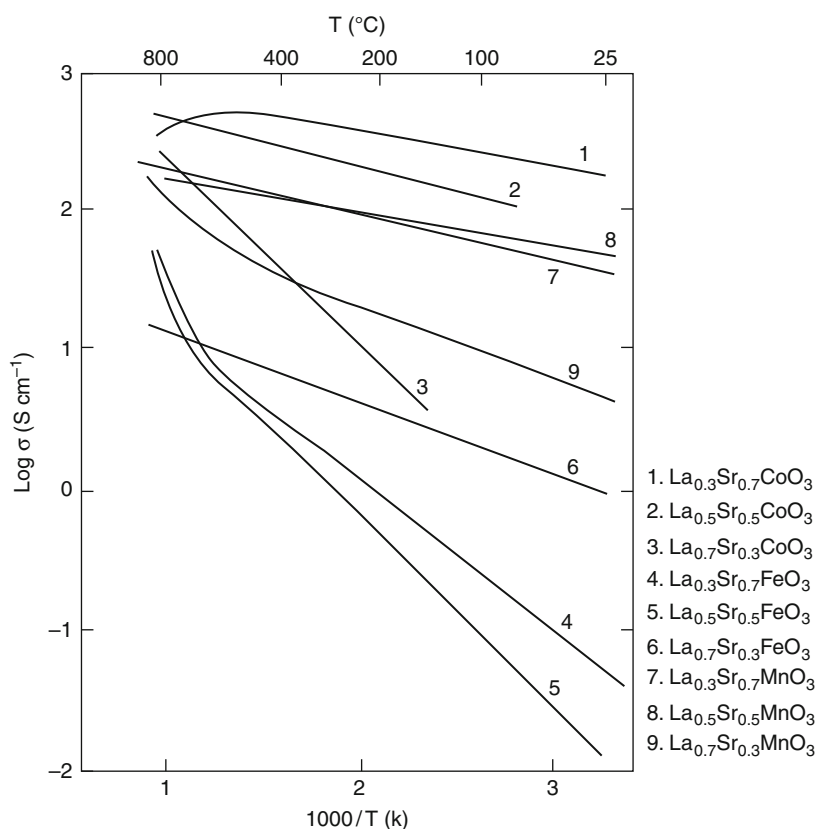


Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 13

Ionic conductivity of proton-conductive perovskite type oxides, reported by Kreuer [156] (Reproduced by permission of Annual Reviews)

conductivity of 1.4×10^4 S/m at $1,000^\circ\text{C}$ [162]. The electronic conductivity increases with increasing alkaline-earth cation (Sr^{2+}) concentration up to $x \approx 0.5$. The increase in Sr content, however, is accompanied with an increase in thermal expansion coefficient [161–168]. The perovskite structure is very tolerant with changes in the ionic radius of the A- and B-site cations, allowing many possibilities to substitute host cations with various dopants. Related perovskite oxides have also been investigated. The A site in this perovskite can be substituted with other rare-earth oxides (e.g., $\text{Y}_{1-x}\text{Ca}_x\text{MnO}_3$) [169–172] and with other alkaline-earth oxides ($\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$) [173–175]. This structure can also accommodate a high concentration of cation vacancies, particularly on the A site as $(\text{La}_{1-x}\text{Sr}_x)_{1-y}\text{MnO}_3$ [22, 176–179]. By substituting

the B site, the electronic conductivity and other properties such as nonstoichiometry, oxygen vacancy concentration, and catalytic activity can be changed considerably [163, 180, 181]. The Mn cation has been substituted with Fe, Co, Ni, and Cu, which enhances the electronic conductivity [180, 182]. Transition metal cations can be substituted on the B site like in $\text{La}_{1-x}\text{Sr}_x\text{Mn}_{1-y}\text{Co}_y\text{O}_3$ or in $\text{La}_{1-x}\text{Sr}_x\text{Fe}_{1-y}\text{Co}_y\text{O}_3$, exhibiting excellent ionic conductivity as well as high electronic conductivity [21, 182–190]. The electronic conductivity of several perovskite oxides is shown in Fig. 14 [180]. Those compounds may be favored for their high electronic conductivity, particularly for planar fuel cell configurations, where in-plane electric current flow is essential in the relatively thin cathode layers. In recent SOFC systems under development,



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 14

Electronic conductivity of SrO -doped LaBO_3 ($B = \text{Mn, Co, Fe}$) after Takeda et al. [180] (Reproduced by permission of ECS-The Electrochemical Society)

Sr -doped LaMnO_3 is still widely used as the state-of-the-art cathode material for ZrO_2 -based electrolytes [2–4, 161, 191].

In addition, cathode materials with mixed ionic and electronic conductivities can expand the electrode reaction area, leading to a decrease in cathodic overvoltage [2, 161, 192]. The perovskite oxides exhibit excellent electronic conductivity, and in specific compounds and conditions, they possess high ionic conductivity. The ionic conductivity of $\text{La}_{0.6}\text{Sr}_{0.4}\text{MnO}_3$ measured by Steele et al. [193] was 10^{-5} S/m at 900°C in air, which is 6 orders of magnitude lower than the ionic conductivity of 8 mol% Y_2O_3 -doped ZrO_2 electrolyte. Therefore, the SrO -doped LaMnO_3 , the state-of-the-art cathode material, is a poor mixed conductor [193]. However, the oxide becomes a mixed conductor in reduced state (at low oxygen partial pressure)

[192, 194], as expected from the defect chemistry [22, 163] and electrocatalytic properties [195].

On the other hand, SrO - or CaO -doped LaCoO_3 and LaFeO_3 are excellent mixed conductors in ambient atmosphere [22, 180, 186–189, 193], where ionic conduction occurs via the oxygen vacancy mechanism [196]. Teraoka et al. [188] have reported that the $\text{La}_{1-x}\text{Sr}_x\text{Co}_{1-y}\text{Fe}_y\text{O}_{3-\delta}$ exhibits ionic conductivity ranging from 10 to 10^3 S/m , while the ionic conductivity of doped ZrO_2 electrolyte is around 10 S/m . High ionic conductivity in $\text{La}_{0.6}\text{Sr}_{0.4}\text{CoO}_3$ has been confirmed by Steele et al. to be 80 S/m at 900°C [193]. Ionic conductivity also increases by substituting the B-site cations with cations with a larger number of 3d electrons from Fe to Co and Ni [182], consistent with the increase in oxygen vacancy concentration [163]. The thermal expansion mismatch between the cathode

Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Table 2 Thermal expansion coefficients of perovskite oxides for SOFC cathode applications [33, 168, 184, 186]

	Materials	Thermal expansion coefficient ($\times 10^{-6}/\text{K}$)	Temp. range ($^{\circ}\text{C}$)
ZrO ₂	ZrO ₂ -8 mol%Y ₂ O ₃	10.5	25–900
LSM	La _{0.85} Sr _{0.15} MnO ₃	11.7	25–900
	La _{0.5} Sr _{0.5} MnO ₃	12.6	25–900
	La _{0.99} MnO ₃	11.2 ^a	25–1,000
	La _{0.94} Sr _{0.05} MnO ₃	11.7 ^a	25–1,000
	La _{0.89} Sr _{0.10} MnO ₃	12.0 ^a	25–1,000
	La _{0.79} Sr _{0.20} MnO ₃	12.4 ^a	25–1,000
	La _{0.69} Sr _{0.30} MnO ₃	12.8 ^a	25–1,000
LSC	La _{0.85} Sr _{0.15} CoO ₃	18.7	25–900
	La _{0.7} Sr _{0.3} CoO ₃	18.3	25–900
	La _{0.5} Sr _{0.5} CoO ₃	22.0	25–900
	La _{0.3} Sr _{0.7} CoO ₃	15.6	25–900
LSCM	La _{0.85} Sr _{0.15} Mn _{0.8} Co _{0.2} O ₃	11.8	25–900
	La _{0.85} Sr _{0.15} Mn _{0.5} Co _{0.5} O ₃	14.6	25–900
	La _{0.5} Sr _{0.5} Mn _{0.5} Co _{0.5} O ₃	14.9	25–900
	La _{0.8} Sr _{0.2} Mn _{0.8} Co _{0.2} O ₃	11.1 ^b	200–800
	La _{0.8} Sr _{0.2} Mn _{0.6} Co _{0.4} O ₃	13.7 ^b	200–800
	La _{0.8} Sr _{0.2} Mn _{0.4} Co _{0.6} O ₃	16.5 ^b	200–800
	La _{0.8} Sr _{0.2} Mn _{0.2} Co _{0.8} O ₃	17.0 ^b	200–800
	La _{0.8} Sr _{0.2} CoO ₃	17.8 ^b	200–800
LSCF	La _{0.8} Sr _{0.2} Co _{0.8} Fe _{0.2} O ₃	18.4	25–900
	La _{0.2} Sr _{0.8} Co _{0.8} Fe _{0.2} O ₃	24.0	25–900
	La _{0.8} Sr _{0.2} CoO ₃	19.7 ^c	100–1,000
	La _{0.8} Sr _{0.2} Co _{0.8} Fe _{0.2} O ₃	20.7 ^c	100–1,000
	La _{0.8} Sr _{0.2} Co _{0.6} Fe _{0.4} O ₃	20.0 ^c	100–1,000
	La _{0.8} Sr _{0.2} Co _{0.4} Fe _{0.6} O ₃	17.7 ^c	100–1,000
	La _{0.8} Sr _{0.2} Co _{0.2} Fe _{0.8} O ₃	15.1 ^c	100–1,000
	La _{0.8} Sr _{0.2} FeO ₃	12.9 ^c	100–1,000

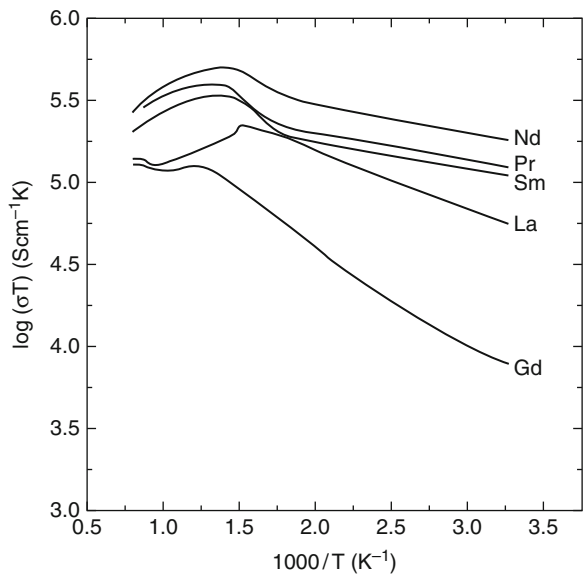
^aSee Ref. [168]

^bSee Ref. [184]

^cSee Ref. [186]

and the electrolyte, as compiled in Table 2, leads to insufficient contact and thus the detachment of cathode layers, which limits the SOFC applications of such mixed conducting cathode materials.

Alternative Cathode Materials Perovskite oxides using smaller lanthanide elements than La, such as Gd_{1-x}Sr_xCoO₃ [197], Sm_{1-x}Sr_xCoO₃ [198], Pr_{1-x}Sr_xMnO₃, and Nd_{1-x}Sr_xMnO₃ [199] have been investigated as cathode materials in order to decrease reactivity with the YSZ electrolyte. Compatibility with YSZ electrolyte was further improved for Gd_{1-x}Sr_xCo_{1-y}Mn_yO₃ [200] by substituting Co with Mn on the B site due to lower reactivity and matching of thermal expansion coefficients. Substitution of Co with Fe on the B site also decreases their thermal expansion coefficient. Perovskite oxides using Co and Fe for the B site, such as Pr_{1-x}Sr_xCo_{1-y}Fe_yO_{3-δ}, Nd_{1-x}Sr_xCo_{1-y}Fe_yO_{3-δ}, and Gd_{1-x}Sr_xCo_{1-y}Fe_yO₃, exhibit high electrical conductivity above 600°C as shown in Fig. 15 and exhibit good catalytic activity on



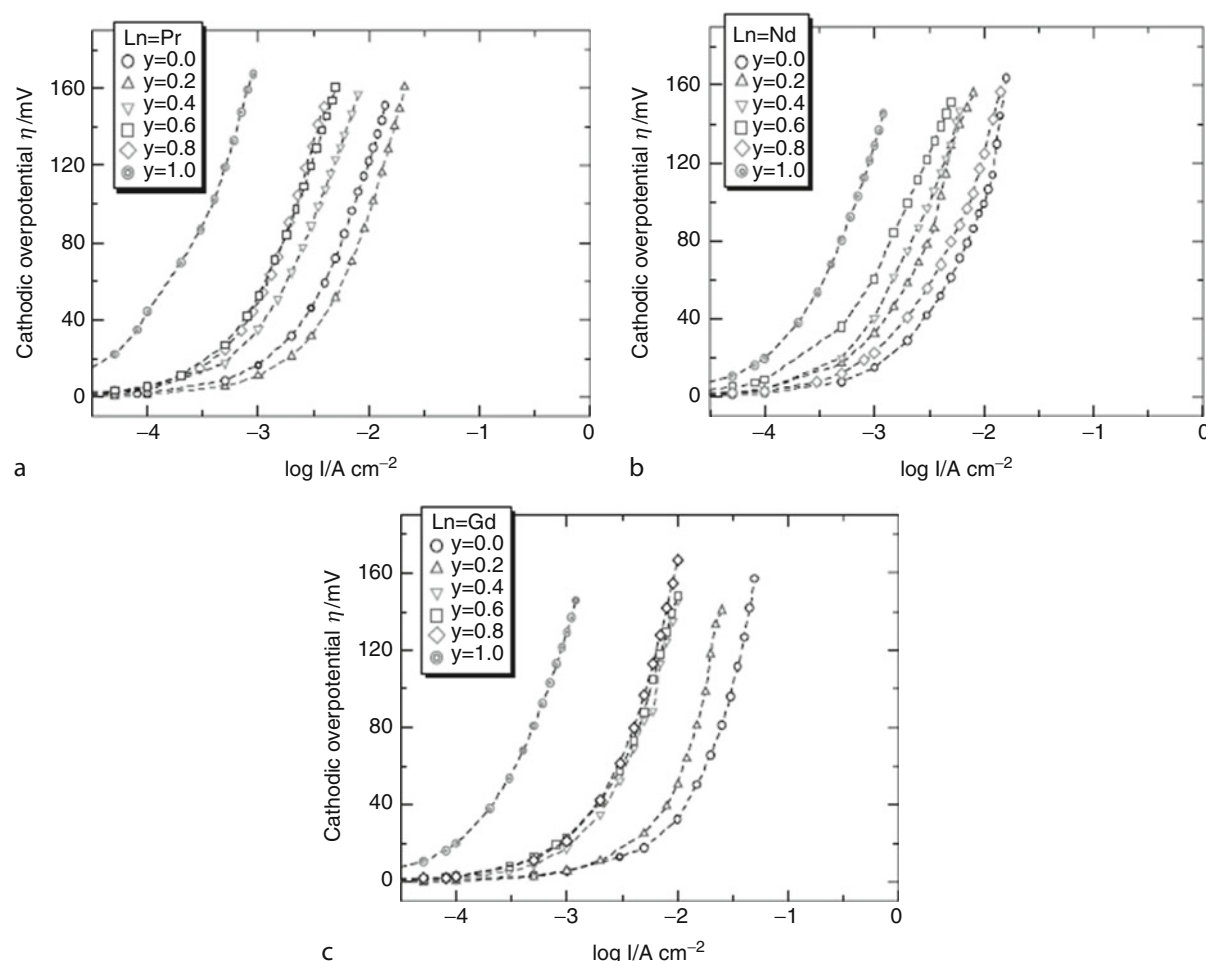
Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 15

Temperature dependence of the electrical conductivity for Ln_{0.4}Sr_{0.6}Co_{0.8}Fe_{0.2}O_{3-δ} (Ln = La, Pr, Nd, Sm, Gd), reported by Tu et al. [201] (Reproduced with permission from Elsevier)

GDC electrolytes in the intermediate operational temperature range (600–800°C) [201, 202], as shown in Fig. 16.

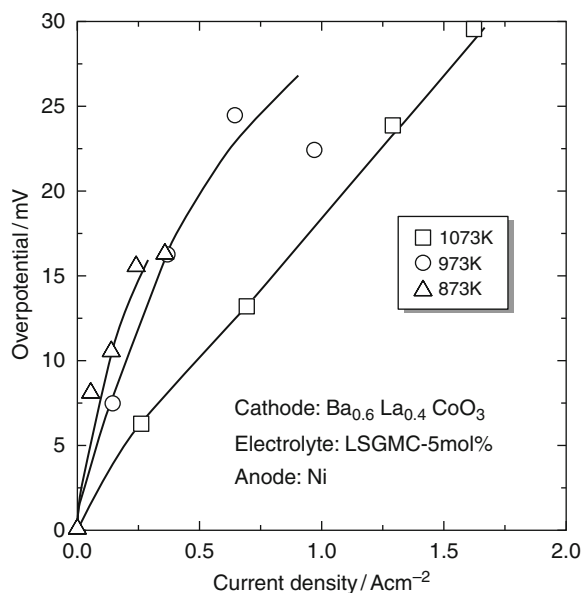
$\text{Sm}_{0.5}\text{Sr}_{0.5}\text{CoO}_3$ has been used as a cathode material for the LSGM electrolyte, and a low cathodic overpotential of less than 50 mV at 0.5 A/cm² was obtained at 873 K [203]. BaCoO_3 was also investigated as a novel cathode material for the LSGM electrolyte. $\text{Ba}_{0.6}\text{La}_{0.4}\text{CoO}_3$ showed a high surface activity for oxygen dissociation and exhibited the same level of cathodic overpotential as $\text{Sm}_{0.5}\text{Sr}_{0.5}\text{CoO}_3$ as shown in Fig. 17 [204].

Perovskite oxides containing Ba in the A site were proposed as cathode materials for a lower-temperature operation. Many of them have been investigated in combination with CeO_2 -based electrolytes. High power density of more than 1 W/cm² was obtained at 600°C by using $\text{Ba}_{0.5}\text{Sr}_{0.5}\text{Co}_{0.8}\text{Fe}_{0.2}\text{O}_{3-\delta}$ (BSCF) for the cathode on a thin SDC electrolyte (12–20 μm in thickness) as shown in Fig. 18 [205]. Enhancement of electrical conductivity by an addition of Sm to BSCF was also reported, which resulted in further improvement of cathode performance at low operational temperatures [206]. $\text{Ba}_{0.5}\text{Sr}_{0.5}\text{Co}_{1-y}\text{Fe}_y\text{O}_{3-\delta}$ possessed low



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 16

Cathodic polarization curves for $\text{Ln}_{0.8}\text{Sr}_{0.2}\text{Co}_{1-y}\text{Fe}_y\text{O}_{3-\delta}$ ((a) $\text{Ln} = \text{Pr}$, (b) $\text{Ln} = \text{Nd}$, (c) $\text{Ln} = \text{Gd}$) cathodes on GDC electrolytes at 700°C, reported by Qiu et al. [202] (Reproduced with permission from Elsevier)

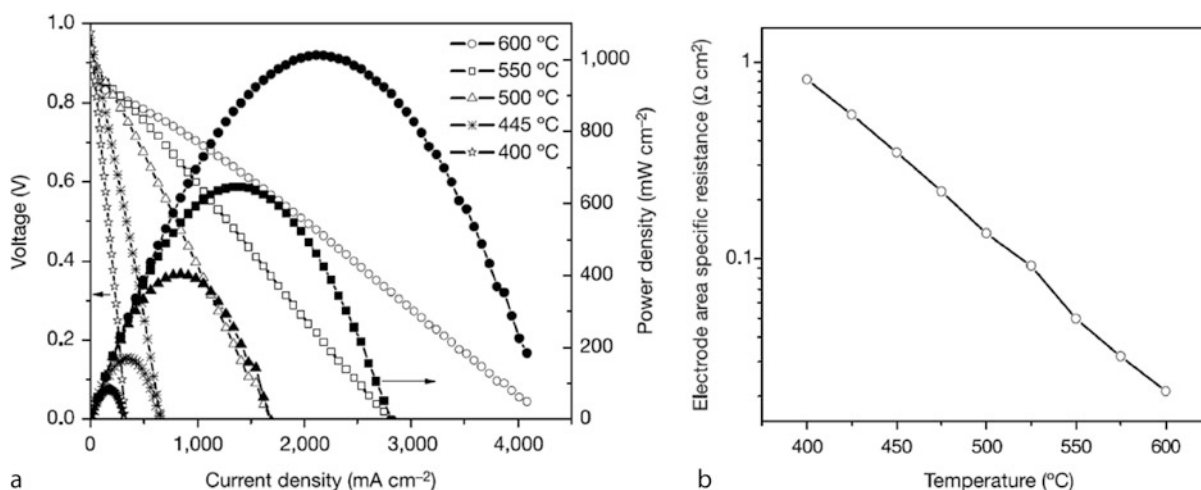


Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 17

Temperature dependence of cathodic overpotential of the Ba_{0.6}La_{0.4}CoO₃ cathode on the LSGMC electrolyte as a function of current density, reported by Ishihara et al. [204] (Reproduced with permission from Elsevier)

vacancy formation energy with low oxygen migration barrier [207]. However, BSCF cathode could degrade in the presence of CO₂ in ambient air, owing to the reaction with alkaline-earth elements, resulting in the formation of Ba and Sr carbonates [208, 209].

Regarding durability of SOFC cathodes, Cr poisoning phenomena of cathode materials have been investigated to better understand long-term cell performance. Taniguchi et al. clarified that not only the vapor pressure of Cr species but also the cathodic overvoltage affects the degradation rate for a cell using La_{0.9}Sr_{0.1}MnO₃ [210]. In addition, reaction between the Cr vapor and the cathode material can decrease the vapor pressure and contribute in delaying the beginning of cell performance degradation [211]. Matsuzaki et al. reported that the use of La_{0.6}Sr_{0.4}Co_{0.2}Fe_{0.8}O₃ (LSCF) on Ce_{0.8}Sm_{0.2}O_{1.9} electrolyte drastically reduced the Cr poisoning in comparison with La_{0.85}Sr_{0.15}MnO₃ [212]. LaNi_{0.6}Fe_{0.4}O₃ (LNF) was also proposed as a novel cathode material, exhibiting high tolerance to the Cr poisoning. Chiba et al. have first reported that LNF possesses suitable properties for operating at low temperatures [213]. LNF has shown high electronic conductivity of



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 18

Performance obtained from a Ba_{0.5}Sr_{0.5}Co_{0.8}Fe_{0.2}O_{3-δ} (~20 μm) | Sm_{0.15}Ce_{0.85}O_{2-δ} (~20 μm) | Ni + Sm_{0.15}Ce_{0.85}O_{2-δ} (~700 μm) fuel cell. (a) Cell voltage and power density as a function of current density and (b) area-specific resistances of the electrodes, reported by Shao and Haile [205] (Reproduced with permission from Nature Publishing Group)

580 S/cm at 800°C and relatively closer thermal expansion coefficient of $11.4 \times 10^{-6} \text{ K}^{-1}$ (30–1,000°C) to that of YSZ than $\text{La}_{0.8}\text{Sr}_{0.2}\text{MnO}_3$ (LSM). Furthermore, it is also revealed that LNF has higher tolerance to Cr poisoning than LSM and LSCF [214]. Figure 19 shows the comparison of cathodic overvoltage by Cr poisoning between LSCF and LNF. Concerning the Cr poisoning phenomena, Jiang et al. [215] explained that the deposition of Cr in the LSM cathode is initiated by the nucleation reaction involving the gaseous Cr species and Mn^{2+} , which are generated by the cathode polarization and regarded as nucleation agents. SrO is identified as a nucleation agent in the case of LSCF. As a result, they concluded that LNF exhibits high tolerance to Cr poisoning because it has no nucleation agent under any electrochemical condition [215]. In terms of electrochemical properties at low temperatures, an electrode resistance of approximately $0.1 \Omega \text{ cm}^2$, corresponding to the cathode polarization, was achieved at 700°C in the case of a LNF-YSZ composite cathode which was made by infiltrating LNF into porous YSZ [216].

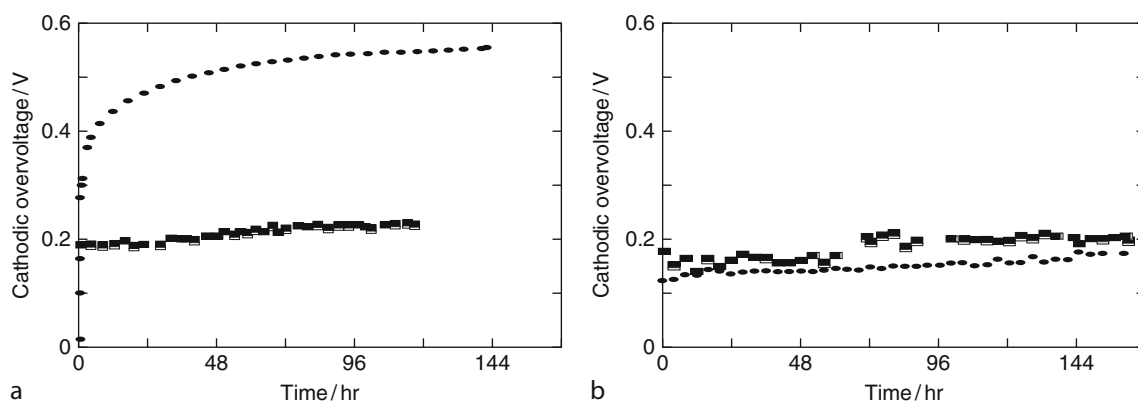
Another type of oxide with the general formulation of $\text{A}_2\text{BO}_{4+\delta}$ (e.g., $\text{La}_{2-x}\text{Sr}_x\text{NiO}_{4+\delta}$ [217], $\text{Nd}_2\text{NiO}_{4+\delta}$ [218, 219], $\text{Pr}_{1.4}\text{Sr}_{0.6}\text{NiO}_4$ [220], and $\text{Sm}_{0.5}\text{Sr}_{1.5}\text{CoO}_{4-\delta}$ [221]) has also been investigated for intermediate-temperature (900–1,000 K) operation. Their K_2NiF_4 -type structure allows for some slight oxygen

overstoichiometry. Oxygen excess, accommodated by oxygen interstitials rather than cation vacancies, leads to high values of oxygen tracer diffusion coefficient. Current density of 0.5 A cm^{-2} at 0.7 V was obtained at 700°C with an anode-supported half-cell with a $\text{Nd}_{1.95}\text{NiO}_{4+\delta}$ cathode and a YSZ electrolyte [219].

In order to enhance the catalytic activity of the cathode for oxygen reduction and thus to increase the electrochemical performance, an addition of noble metals has also been investigated [222]. By the addition of silver to a LSCF- CeO_2 composite cathode, an improvement of cathode performance at lower temperatures below 700°C was measured [223, 224]. Consequently, Ag improved the surface exchange step of oxygen from the gas phase into the oxide [223]. Cell performance of more than 500 mW/cm^2 at 0.7 V at 700°C was obtained using anode-supported cells with LSCF-coated Ag cathode and YSZ electrolyte [225]. Generally speaking, Sr-doped LaMnO_3 and related materials are suitable for high-temperature operation of SOFCs, while more electrochemically/catalytically active alternative cathode materials may be needed especially for lower operational temperatures.

Anode Materials

Current Anode Materials In state-of-the-art SOFCs, Ni-based cermet anodes are commonly used, which



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 19

Cathodic overvoltage (IR corrected) for cells using (a) LSCF cathode and (b) LNF cathode as a function of time: (●) cell with Inconel 600 at 0.7 A cm^{-2} and (■) cell without Inconel 600 at 0.7 A cm^{-2} , reported by Komatsu et al. [214] (Reproduced by permission of Elsevier)

exhibit excellent electrocatalytic activity for anodic electrochemical reactions, fuel oxidation, and internal reforming, as well as sufficient current collection [2–4, 161, 226]. Due to the requirements of operation in reducing environments and high electronic conductivity, pure metallic electrodes are generally used as anodes. Several metals such as Ni, Ru, Pt, and Co have been studied as anode materials [161, 227, 228]. Ni exhibited the lowest anodic overpotentials [161, 227], while Ru could be an excellent anode material [228]. Pure Ni has electronic conductivities of $1.4 \times 10^5 \text{ Scm}^{-1}$ and $2.6 \times 10^4 \text{ Scm}^{-1}$ at 298 K and 900 K, respectively; a thermal expansion coefficient of ca. $14.6 \times 10^{-6} \text{ K}^{-1}$; and a melting point of $1,455^\circ\text{C}$ [229]. The transition metals, which show higher activities for the oxidation of hydrogen, tend to change their microstructures at high temperatures under the reducing conditions of the anode, that is, coalescence of metal particles leads to anode densification. To stabilize the anode morphology in terms of porosity and thermal expansion coefficients, in general, cermets are used, where the ceramic component provides a rigid network and hinders the porous anode morphology from coalescence. The ceramic component also ensures to make the thermal expansion coefficient similar to that of the electrolyte materials.

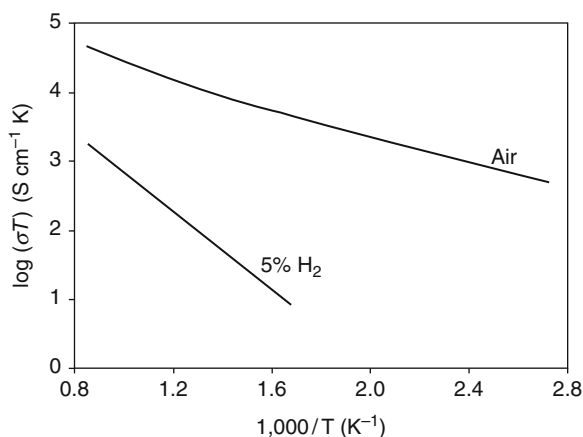
Internal reforming of hydrocarbon-based fuels can also occur on the Ni surface in the cermet anode. In an SOFC cermet anode, substantially long triple phase boundaries (TPB) can be achieved, and given that solid networks of two phases (electronic conductors and ionic conductors) are obtained throughout the anode volume. It has been well accepted that the electrochemical reaction occurs around the TPB, where oxygen ion conductor (electrolyte), electron-conducting metal phase (Ni), and gas phase all meet together [2, 161].

Alternative Anode Materials

General While Ni/YSZ cermet is the most preferred anode material for SOFCs operating with hydrogen-containing fuels, there are several efforts to find alternative materials for the SOFC anode. Many studies on oxide anodes have been made for perovskite-related oxides. LaCrO_3 [230–237], LaFeO_3 [238], LaTiO_3 [239], and SrTiO_3 [240–243] systems are considered as possible alternative anode materials.

Perovskite-Type Oxide LaCrO_3 -based oxides are well known as interconnect materials for SOFCs. They have high durability in reducing atmosphere. The oxygen-deficient perovskite $\text{La}_{0.75}\text{Sr}_{0.25}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_3$, used as an anode material by Tao et al. [230], is stable in both fuel and air atmospheres and has revealed stable electrode performance, as shown in Fig. 20 [230]. However, this material has a low electronic conductivity in reducing anode atmosphere and is not stable to sulfur impurities in the fuel [231]. Its long-term performance has been proven unsatisfactory [231]. $\text{Sr}_2\text{MgM}_0\text{O}_{6-8}$ exhibited a Sulfur-tolerance up to 50 ppm H_2S [232]. Experimental results on the $\text{La}_{0.8}\text{Sr}_{0.2}\text{Cr}_{0.97}\text{V}_{0.03}\text{O}_3$ (LSCV)-YSZ composite anode have demonstrated that the electrochemical performance is comparable to that of Ni/YSZ after 48 h operation [233].

Ruiz-Morales et al. [234] demonstrated symmetrical fuel cells (SFCs) using the same material $\text{La}_{0.75}\text{Sr}_{0.25}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_{3-\delta}$ (LSCM) on both the anode side and the cathode side. Due to its enhanced electrochemical properties under both reducing and oxidizing conditions, LSCM-based SFCs offered sufficient performances, 0.5 and 0.3 Wcm^{-1} at 950°C using H_2 and CH_4 as fuels, respectively [234].



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 20

Temperature (T) dependence of total conductivity (σ) of $\text{La}_{0.75}\text{Sr}_{0.25}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_{3-\delta}$ in air and 5% H_2 , reported by Tao and Irvine [230] (Reproduced with permission from Nature Publishing Group)

Sauvet et al. [235] have studied the electrochemical properties of $\text{La}_{1-x}\text{Sr}_x\text{Cr}_{1-y}\text{Ru}_y\text{O}_{3-\delta}$ ($x = 0.2-0.4$, and $y = 0.02$ and 0.05) for hydrogen and methane fuels at 750 and 850°C. They found the best performances obtained with 30% of Sr doping, and the variation of the ruthenium content, approximately between 0.02 and 0.05, had little influence on the electrochemical results. At 750°C, the polarization resistance was $3.7 \Omega \text{ cm}^2$ in hydrogen and $40 \Omega \text{ cm}^2$ in methane for the YSZ/CeO₂/graded $\text{La}_{0.7}\text{Sr}_{0.3}\text{Cr}_{0.95}\text{Ru}_{0.05}\text{O}_{3-\delta}$ -YSZ structure [235]. Sauvet et al. also studied $\text{La}_{1-x}\text{Sr}_x\text{Cr}_{1-y}\text{Ni}_y\text{O}_{3-\delta}$ compositions as novel anode materials. They have found that these materials have catalytic activity, but less than that of Ni composites [236, 237].

Lanthanum ferrites have also been investigated as potential anode materials. They are already known for high mixed electronic and ionic conductivity and good oxygen exchange kinetics. However, they have poor stability at lower oxygen partial pressures in the reducing atmosphere. The stability, however, could be improved by mixing with GDC [238].

(La, Sr)TiO₃ has been investigated as alternative SOFC anodes, along with other A-site-deficient lanthanides. Nevertheless, it has poor conductivity than Ni/YSZ cermet [239]. Doped SrTiO₃ could be an interesting anode material, and the mixed conductivity of La-doped and Fe-doped SrTiO₃ is widely influenced by the chemical composition [240, 244, 245]. SrTiO₃ doped with Y [241] and La [242] exhibits increased electrical conductivity. Unfortunately, the high-temperature electrical conductivity of various perovskite anodes is still lower than that of the conventional Ni/YSZ cermet for the use as SOFC anodes [243].

Other Transition Metal Cermets The Cu-based anode cermet has been extensively studied. For example, Park et al. [246, 247] demonstrated that the anode composed of $\text{Cu}/\text{Sm}_2\text{O}_3\text{-CeO}_2/\text{YSZ}$ was active for direct electrochemical oxidation of alkanes and alkenes, such as methane, ethane, butane, butene, hexane, and toluene, at the operating temperature of 700°C. CeO₂ plays an important role as a primary catalyst material for oxidizing hydrocarbons, and the main role of Cu is to provide electrical conductivity. Zhou et al. [248] reported good performance of

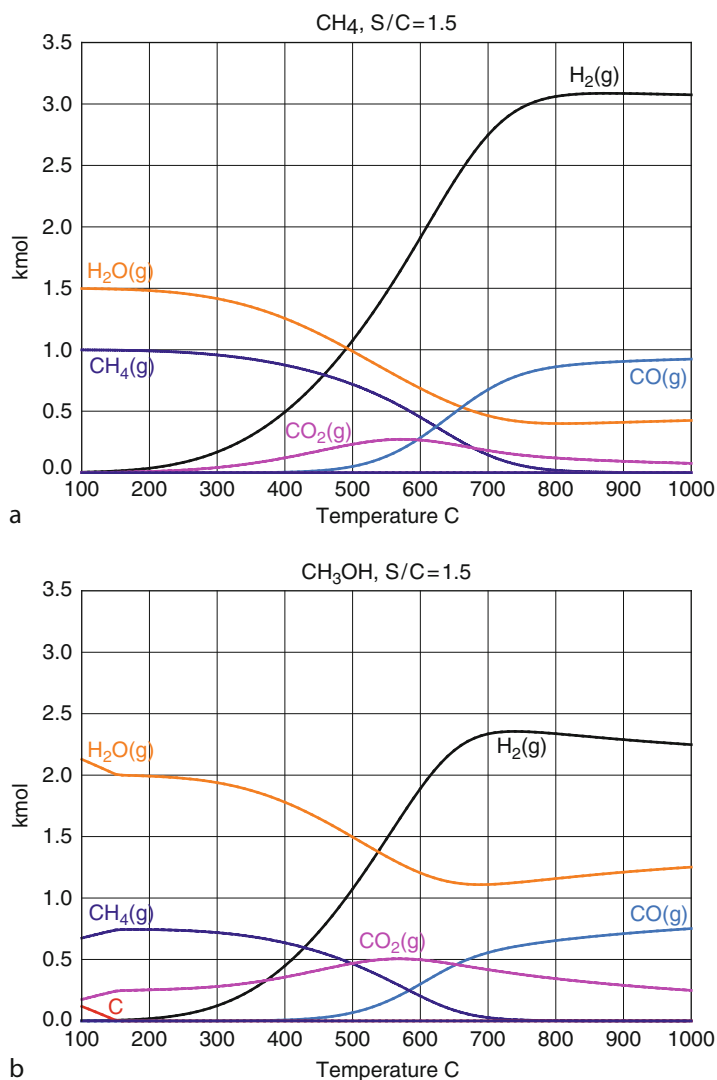
CeO₂-Rh anodes for the direct electrochemical oxidation of waste vegetable oil, which led to stable current density of 500 mA cm^{-2} over 100 h under the condition of 0.5 V and 700°C.

Alternative Fuels

Fuel Flexibility

SOFCs are the most flexible fuel cells with respect to the type of fuels [1, 246–255] which can be supplied directly to the anodes. This is partly due to their high operational temperatures suitable for internal reforming within the anodes. Because of their multi-fuel capability, not only hydrogen and carbon monoxide [256] but also various types of fuels such as hydrocarbons, alcohols, and biogas could be used via internal reforming and/or via simple external reforming [246–249, 257–260]. Practical fuels comprise natural gas (consisting mainly of CH₄ with a small amount of other hydrocarbons such as C₂H₆), coal gas (consisting mainly of CO and H₂), liquefied petroleum gas (LPG, consisting mainly of C₃H₈ with C₄H₁₀), naphtha (consisting mainly of C₅ and C₆ hydrocarbons), gasoline (consisting mainly of hydrocarbons with carbon numbers around 8), kerosene (consisting mainly of hydrocarbons with carbon numbers around 12), diesel fuel (consisting mainly of hydrocarbons with carbon numbers around 14–16), alcohols, dimethyl ether, and biogas (consisting mainly of CH₄, CO₂, and H₂O) [251, 252].

In any case, one of the most important issues to be prevented in SOFC systems is carbon deposition (coke formation) from the fuels. Figure 21 shows the equilibrium products for (a) methane- and (b) methanol-based fuels with the steam-to-carbon (S/C) ratio of 1.5 at elevated temperatures [251]. Assuming thermochemical equilibrium, carbon deposition is not expected to occur within a wide temperature range. The calculated results for various other fuels mentioned above have been shown elsewhere [251]. The minimum amounts of H₂O (water vapor) necessary to prevent carbon deposition are shown in Fig. 22 for hydrocarbon fuels. While S/C of 1.5 is enough for CH₄, higher S/C is needed with increasing carbon number of hydrocarbons, especially at lower



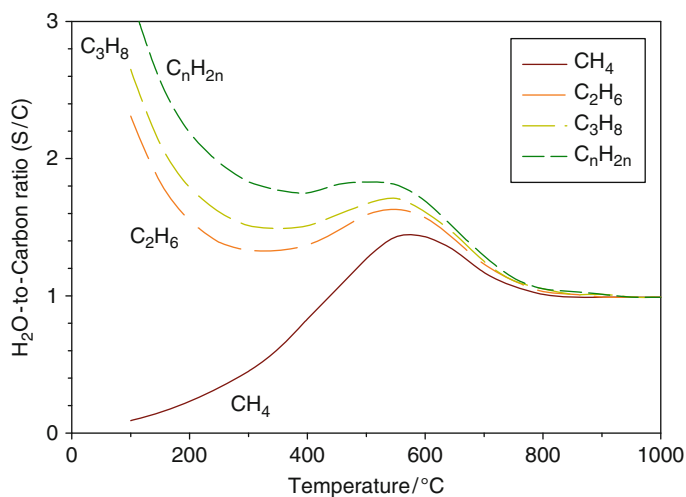
Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 21

Equilibrium products from (a) methane- and (b) methanol-based fuels with the steam-to-carbon ratio of 1.5

temperatures. Such dependencies have also been revealed for O₂ (partial oxidation) and CO₂ (CO₂ reforming) [251] to prevent carbon deposition.

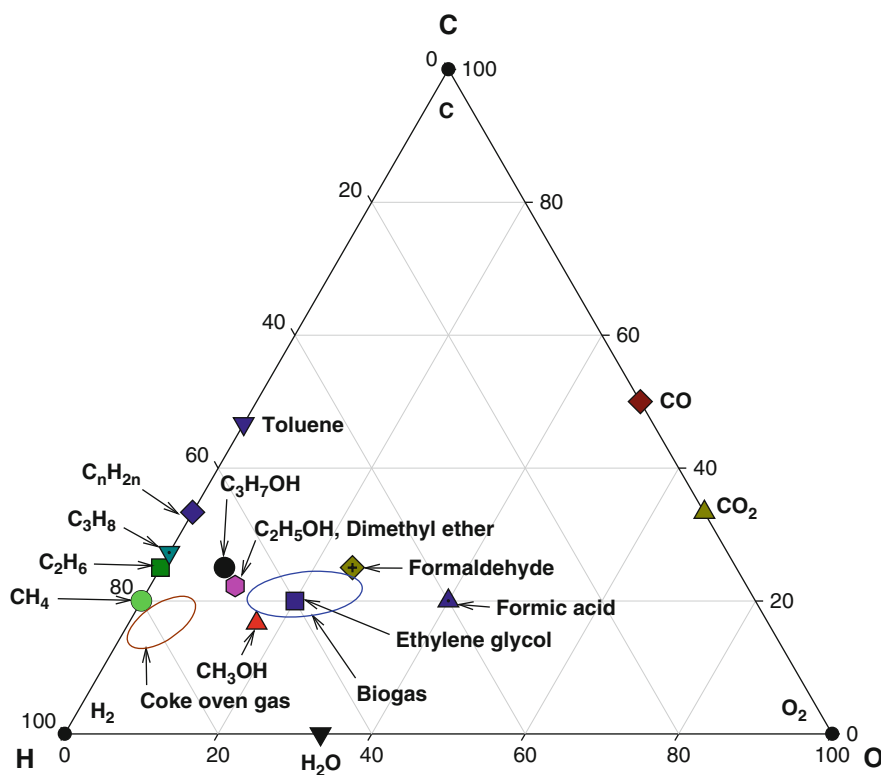
It has been found that the major equilibrium constituents in fuel gases are H₂(g), H₂O(g), CO(g), CO₂(g), CH₄(g), and C(s) [251]. Since their equilibrium compositions depend solely on the C–H–O ratio, we can plot, on C–H–O diagrams [252, 253, 255], parameters relevant to optimize operational conditions, including carbon deposition region, gas partial pressures, and electromotive force. Figure 23 shows the positions of various fuel species for fuel cells [252].

Their positions can be shifted by adding H₂O (steam reforming), O₂ (partial oxidation), and/or CO₂ (CO₂ reforming) as co-reactants. Figure 24 shows the carbon deposition limit lines, the carbon-rich side of which corresponds to the carbon deposition region at each temperature. From Figs. 23 and 24, we see that an addition of co-reactants is necessary to prevent carbon formation at SOFC operational temperatures from most fuels shown in Fig. 23. Once the C–H–O ratio is specified, C–H–O ternary diagrams are useful to derive relevant operational parameters without any additional thermochemical calculations [252].



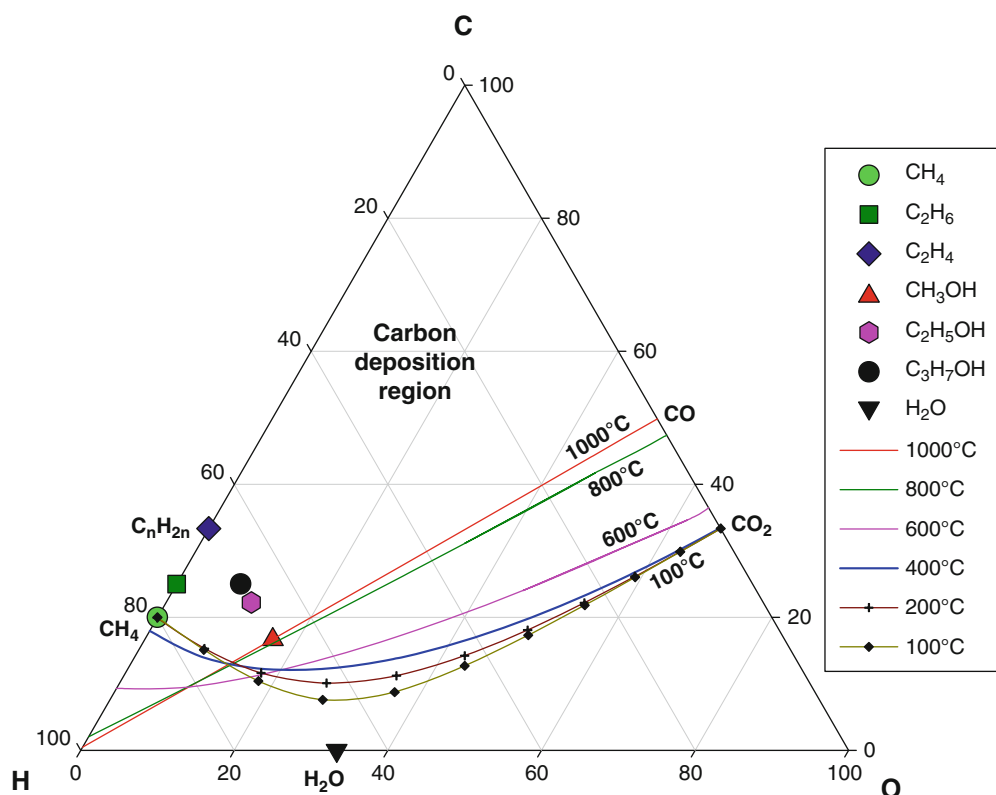
Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 22

Minimum steam-to-carbon (S/C) ratio needed to prevent carbon deposition in thermodynamic equilibrium for hydrocarbons [251] (Reproduced with permission from ECS – The Electrochemical Society)



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 23

Various hydrocarbons and related species in the C–H–O diagram [252] (Reproduced with permission from ECS – The Electrochemical Society)



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 24

Carbon deposition limit lines at various temperatures in the C–H–O diagram [252] (Reproduced with permission from ECS – The Electrochemical Society)

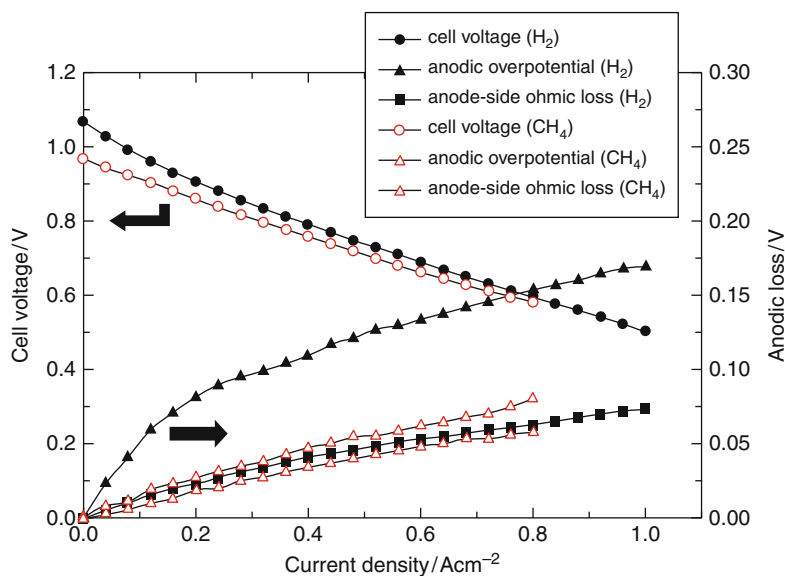
Apart from such thermochemical processes, also, kinetic phenomena may dominate electrode performance. In the following sections, various practical fuels are experimentally examined, the use of which may enable broader application of SOFCs.

Practical Fuels

City gas (town gas) is used as the state-of-the-art fuel for the first-generation commercial SOFC systems. City gas, consisting mainly of CH_4 from the natural gas, may be partially pre-reformed before supplying to the SOFC systems. Figure 25 shows the comparison of the current-voltage (I – V) characteristics between a H_2 -based fuel and a typical SOFC fuel (50% pre-reformed CH_4 with $\text{S/C} = 2.5$) [261]. The cross section of a typical SOFC used for this electrochemical measurement (see Fig. 25) is shown in Fig. 26. This simulated practical SOFC fuel for distributed cogeneration

systems consists of the one-to-one mixture of fully reformed CH_4 with $\text{S/C} = 2.5$ and non-reformed CH_4 with $\text{S/C} = 2.5$. While too low S/C may result in carbon deposition, excessive amount of H_2O causes lower OCV and may cause re-oxidation of Ni at high fuel utilization of the SOFC systems.

As practical SOFC fuels may contain a small amount of sulfur as an impurity, H_2S poisoning should be considered with respect to various operational conditions [262–266]. As an example, Fig. 27 shows the cell voltage, anodic polarization, and the ohmic loss on the anode side, measured for 50% pre-reformed CH_4 -based fuel with $\text{S/C} = 2.5$ containing 5 ppm H_2S . An initial cell voltage drop followed by a quasi steady-state cell voltage can be observed associated mainly with an increase in anodic polarization and almost constant IR loss, up to 3,000 h [262]. Exactly speaking, cell voltage generally decreased up to ca. 200 h and then remained almost constant but with a slightly higher degradation



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 25

Comparison of the cell voltage and anodic losses in using 3% humidified H_2 and 50% pre-reformed CH_4 ($\text{S/C} = 2.5$) as fuels at 800°C . Anodic loss includes anodic overpotential and anode-side ohmic loss [261] (Reproduced with permission from Dr. Kengo Haga)

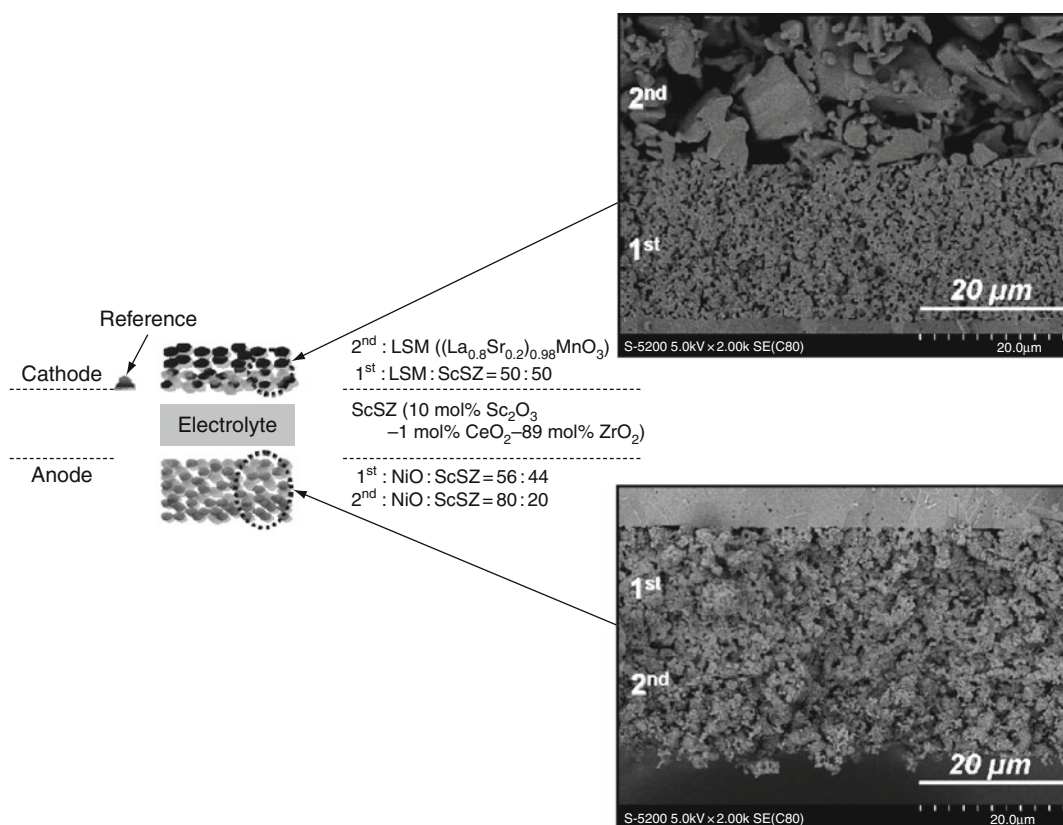
rate of 0.68%/1,000 h with 5 ppm H_2S , compared to the degradation rate of 0.3%/1,000 h without H_2S impurity [262]. This result clearly shows that SOFC is more tolerant against sulfur poisoning compared to other types of fuel cells, giving more flexibility in practical fuel selection. Even so, it should be noted that carbon deposition phenomena can be affected by the simultaneous presence of sulfur impurities and non-reformed hydrocarbons [262].

Poisoning mechanisms of sulfur, a typical common fuel impurity, have been extensively analyzed for various fuels, including H_2 , $\text{H}_2\text{-CO}$, CH_4 , partially reformed CH_4 , and simulated biogas [263–266]. For relatively low concentrations (ppm level) of sulfur, reversible processes associated with adsorption/desorption of sulfur have been considered as the predominant mechanism [263–265]. At higher sulfur concentration and/or lower operational temperature, an irreversible degradation, associated with the oxidation of Ni due to a large anodic polarization, has been observed [263, 264]. Sulfur poisoning to internal reforming reactions is also serious, so that much larger cell voltage drop has been observed in the case where CH_4 -rich fuels are supplied [265]. In addition, the

C–H–O–Ni–S stability diagram suggests that the formation of Ni_3S_2 (melting point 787°C) could be possible for hydrogen-poor fuels. Possible sulfur poisoning mechanisms are summarized elsewhere [262, 265].

Poisoning phenomena have been analyzed for other impurities, including chlorine, siloxane, phosphorus, and boron on the anode side. Figure 28 shows FESEM micrographs of anodes poisoned by these impurities of ppm levels [266]. Chlorine poisoning is associated with the sublimation of NiCl_2 [267], while siloxane poisoning leads to silica deposition [266]. Phosphorus is reactive with Ni to form a eutectic compound [268]. The presence of boron accelerates the grain growth of Ni in the anodes [266].

Various other types of fuels have been examined as possible SOFC fuels. Direct feeding of hydrocarbon fuels has been reported including gaseous fuels such as methane [246, 247, 269], biogas [270], ethane [246, 247], and butane [246, 247], and liquid fuels such as n-decane [271], gasoline [271], synthetic diesel [272], crude and jet fuel oils [273], vegetable oil [248], and biodiesel [274]. Generally speaking, the S/C ratio should be optimized to prevent carbon deposition for



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 26

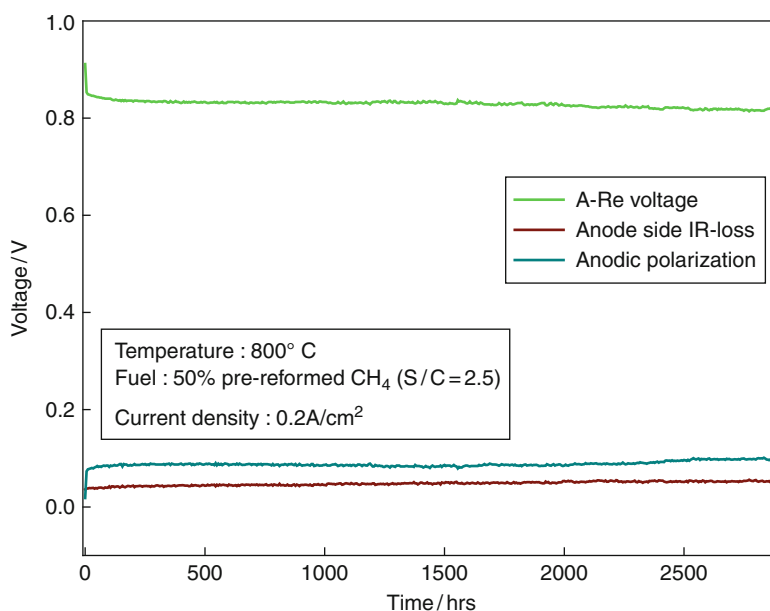
FESEM micrographs of typical cathode and anode. Electrochemical performance of this cell is shown in Fig. 25. The composition of each electrode layer is also shown [261] (Reproduced with permission from Dr. Kengo Haga)

internal steam reforming, and the oxygen-to-carbon (O/C) ratio should be adjusted for internal partial oxidation. Fuel impurities in such practical fuels should also be taken into account to maintain cell performance and durability.

Alcohols

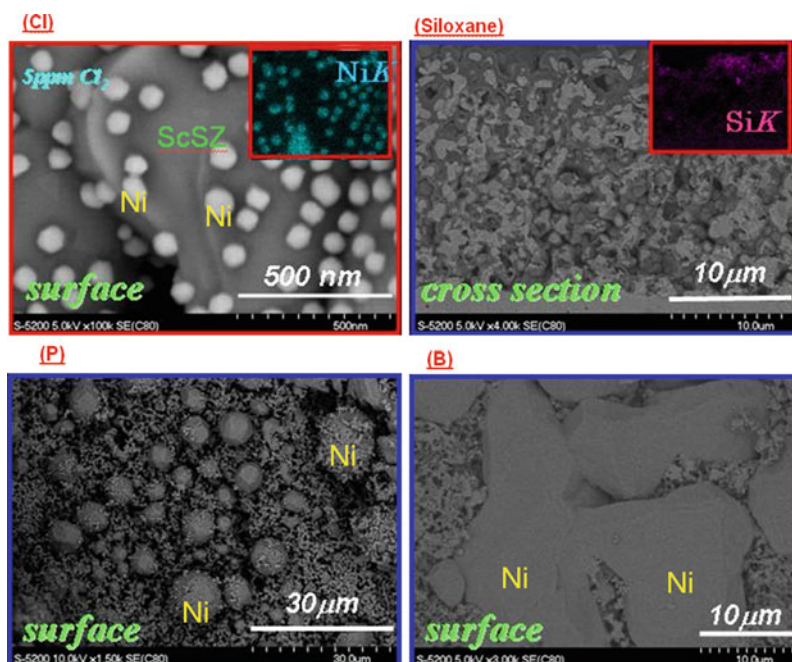
The use of alcohols as SOFC fuels is of technological interest, as alcohols (bioliquids) can be derived from biomass and can be easily reformed. In addition, direct-alcohol SOFCs could be an alternative to direct-methanol fuel cells, typically using polymer electrolytes at a low temperature [4]. Power generation characteristics for alcohol-based fuels and for the simulated reformed gas (consisting mainly of H_2 , CO, and H_2O) are shown in Fig. 29 [257]. All these alcohols were premixed with water, so that the C–H–O ratio was

fixed to the same value (C:H:O = 9:38:10 for all fuels). It has been demonstrated that direct-alcohol SOFCs are feasible, at least, for alcohols with carbon number up to 4 (butanol) [257, 258]. In case that methanol was directly supplied, the I–V characteristics were similar to those for the simulated reformed gas. However, with increasing carbon number of alcohols, a decrease in cell voltage was observed. From the gas analysis by gas chromatograph, it has been revealed that the compositions of the simulated reformed gas and the methanol-derived fuel gas were almost the same, well explaining the similar I–V characteristics [257]. However, with increasing carbon number of alcohols, a decrease in H_2 and CO concentrations associated with a decrease in cell voltage [257] was observed. Therefore, a higher catalytic activity to reform alcohols at fuel electrodes (or pre-reforming) is necessary to optimize the performance of such direct-alcohol SOFCs.



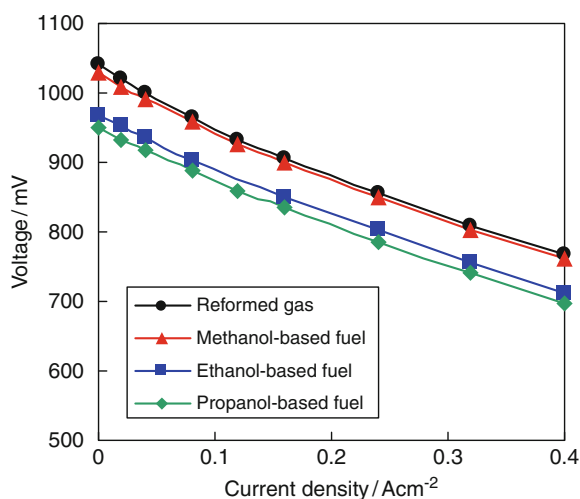
Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 27

Cell voltage, anodic polarization, and anode-side IR loss, measured at 200 mA/cm² during H₂S (5 ppm) poisoning at 800°C for the 50% pre-reformed CH₄ fuel with S/C = 2.5



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 28

Microstructural change of anodes by various impurities: chlorine, siloxane, phosphorus, and boron. Nanocrystalline Ni precipitates after chlorine poisoning, melted Ni particles after phosphorus poisoning, and grown Ni grains after boron poisoning can be distinguished, while deposited SiO₂ (dark grains) filling the porous Ni-cermet (bright grain) anode can be seen in the cross section of the anode after the siloxane poisoning [262, 266] (Reproduced with permission from Elsevier)



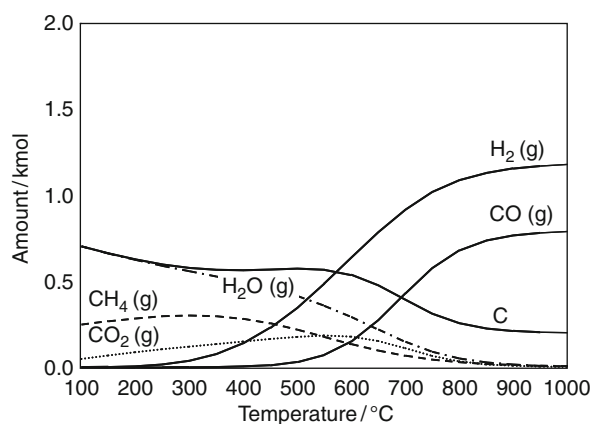
Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 29

Current-voltage characteristics of SOFCs for alcohol-based fuels at 1,000°C [257] (Reproduced by permission of ECS-The Electrochemical Society)

Biogas

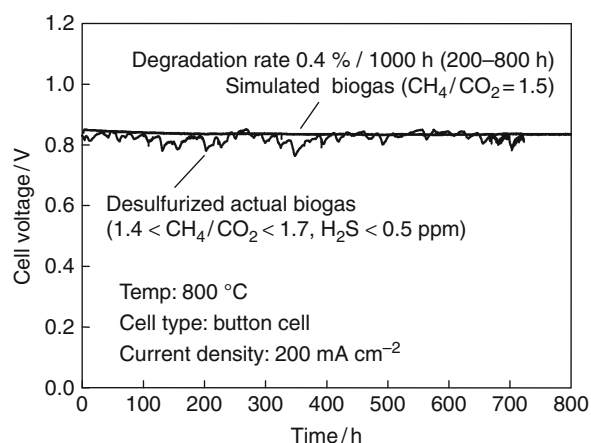
The use of biogas as a fuel for SOFCs is of both technological and environmental interest [3, 4, 251, 260, 270, 275–278]. Biogas can be produced from biomass via various procedures such as fermentation and thermal decomposition. Organic municipal wastes (biowastes) are all metabolized by bacteria under anaerobic conditions producing a mixed gas consisting of ca. 60% CH₄ and 40% CO₂. Anaerobic fermentation which proceeds in a reducing atmosphere enables us to extract chemical energy of biowastes as CH₄-rich biogas. Figure 30 shows the equilibrium products from biogas initially consisting of CH₄ and CO₂. High-temperature equilibrium gas from the biogas mainly consists of H₂ and CO, so that a comparable electrochemical performance with hydrogen-based fuels [275, 277] is expected. This figure also indicates that carbon formation is thermochemically expected, so that an addition of H₂O (water vapor) and/or O₂ is needed.

As practical examples, Fig. 31 shows the cell voltages of internal reforming SOFC running on biogas measured at 200 mA cm⁻² [278]. Simulated biogas with the CH₄/CO₂ ratio of 1.5 had led to stable cell voltage above 0.8 V for 800 h without pre-reformer and humidifier. The degradation rate of only 0.4%/1,000 h



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 30

Equilibrium products from biogas (CH₄ 60% and CO₂ 40%). The calculations are performed by assuming that 1 kmol carbon is supplied



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 31

Performance of direct-biogas SOFC: 1 month test of internal reforming SOFC running on simulated and desulfurized actual biogases at 800°C using anode-supported button cells

has proven that the biogas-fueled SOFC can be operated in the internal reforming mode. As for the actual biogas, a rather high voltage comparable to that obtained by simulated biogas was achieved for 1 month. Monitoring of biogas composition simultaneously with the cell voltage has showed that voltage

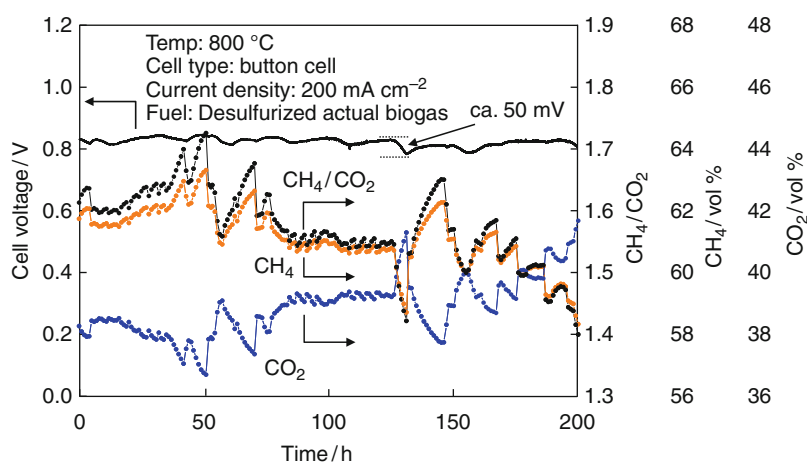
fluctuation (a maximum of 50 mV level) is related to the fluctuation of the CH_4/CO_2 ratio [270] (Fig. 32). Abrupt decrease in CH_4 concentration (or abrupt increase in CO_2 concentration) induced temporarily lower cell voltage. The CH_4/CO_2 ratio in the actual biogas fluctuated between 1.4 and 1.7, corresponding to a CH_4 concentration ranging between 58 and 63 vol%. Biogas composition is influenced by many factors, for example, kinds of organic wastes, their physical states, and operational conditions of methane fermentation such as temperature and pH of the waste slurry.

It should be noted that carbon deposition was more noticeable for the desulfurized actual biogas compared to simulated biogas. Judging from the result of the gas analysis, carbon deposition is probably triggered by traces of H_2S in the desulfurized biogas which can cause deactivation of Ni catalyst for the dry reforming of methane [270]. Acceleration of carbon deposition in the presence of trace H_2S was also reported for SOFCs operated with partially reformed CH_4 -based fuels [262]. Air addition to biogas is effective in preventing carbon deposition [277]. Solid oxide fuel cells could therefore become a core technology in a carbon-neutral zero-emission renewable energy systems, as biogas can be efficiently converted to electricity and heat, while exhaust CO_2 gas may be reused in photosynthesis of biomass using solar energy.

Coal Gas

As one of the most abundant and inexpensive fossil energy resources [279], coal can be used to produce coal gas, with CO and H_2 as main constituents. Coal gas could be applied as a fuel for a combined-cycle power plant with coal gasifier, SOFCs, and gas (and steam) turbine(s) [2, 256, 280–284]. The concentration of water vapor and carrier gas in such a gas depends on the gasification procedures for the partial oxidation of coal by air, oxygen, or water [280]. It is therefore of technological relevance to analyze power generation characteristics of SOFCs operated with mixed H_2 and CO fuel gases, with consideration that coal gas is CO-rich in general.

Figure 33 shows the I-V characteristics of a SOFC at $1,000^\circ\text{C}$ for different H_2 -to-CO ratios [256]. The open-circuit voltage increased slightly with decreasing fraction of CO gas, expected from the oxygen partial pressures in thermochemical equilibrium. Figure 33 shows that the I-V characteristics improved, that is, the cell voltage increased for a given current density with decreasing CO concentration of the mixed gas. This change in the I-V characteristics became significant with increasing current density, but this result clearly indicates that the mixture of CO and H_2 , the major constituents of coal gas, can be applied as SOFC fuels, provided that fuel impurity effects can be avoided.



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 32

Voltage fluctuation in synchronization with the fluctuation of biogas composition during the long-term test (initial 200 h of Fig. 31 is shown)

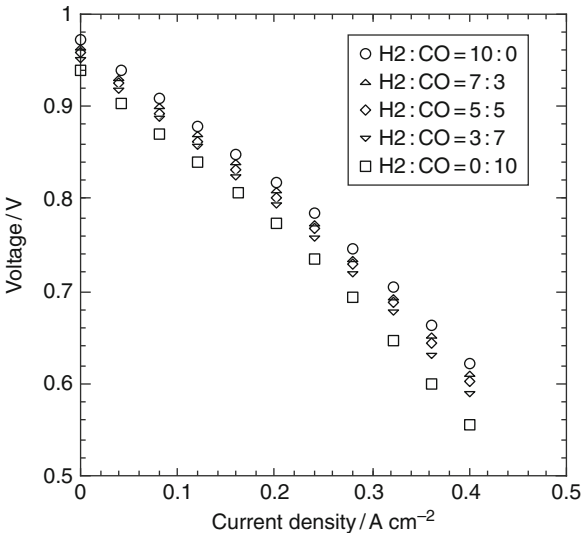
It should be noted that raw coal gas contains various kinds of fuel impurities, so that poisoning effects should be carefully examined. Influence of typical fuel impurities in coal gas on electrochemical performance and

microstructural stability has been extensively studied and discussed in the literature [262–266, 268, 285–290].

Future Directions

The first-generation SOFC systems just commercialized are constructed using the most common materials such as stabilized ZrO₂ electrolytes, La(Sr)MnO₃ cathodes, and Ni cermets, and applying the most conventional fuels such as city gas. However alternative materials and fuels could expand the future applicability of SOFCs. Table 3 briefly summarizes the state-of-the-art and alternative materials for electrolytes, cathodes, and anodes, as well as possible fuels and applications. Advanced materials may offer alternatives too, for example, high-temperature polymer electrolyte fuel cells, direct-alcohol fuel cells, and portable fuel cells, as well as fuel cells for power generation, cogeneration, and transportation. Advanced fuel electrode materials may realize flexible SOFCs applicable to wide varieties of fuels. Nanostructural design may considerably improve their electrochemical performance, for example, by tailoring nano-size and stress effects, by reducing grain boundary resistivity, as well as by introducing fast ionic/mixed conducting pathways.

While most of the alternative materials and fuels mentioned in this entry have not yet been tested in fuel cell stacks/systems or even as a single cell, they may



Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Figure 33 Typical I-V characteristics of SOFCs for different H₂-to-CO ratios at 1,000°C [256] (Reproduced with permission from ECS – The Electrochemical Society)

Fuel Cells (SOFC): Alternative Approaches (Electrolytes, Electrodes, Fuels). Table 3 State-of-the art technologies and typical alternatives of SOFC materials, fuels, and applications

	Electrolyte materials	Cathode materials	Anode materials	Fuels	Possible applications
State-of-the-art technologies	ZrO ₂ (Y ₂ O ₃) (YSZ) ZrO ₂ (Sc ₂ O ₃) (ScSZ) La(Sr)Ga(Mg)O ₃	La(Sr)MnO ₃ La(Sr)Co(Fe)O ₃ Sm(Sr)CoO ₃	Ni/YSZ Ni/ScSZ Ni/CeO ₂ (Gd ₂ O ₃)	City gas, LP gas, H ₂	Stational cogenerations, Distributed power units, Combined power plants (SOFC/turbine)
Typical alternatives	CeO ₂ (Gd ₂ O ₃) Gd ₂ (Zr _x Ti _{1-x}) ₂ O ₇ Ba ₂ In ₂ O ₅ La _{9.33} Si ₆ O ₂₆ La _{9.33} Ge ₆ O ₂₆ BaCe(Y)O ₃ BaZr(Y)O ₃	Ba(Sr)Co(Fe)O ₃ Gd(Sr)CoO ₃ Sm(Sr)CoO ₃ La ₂ (Sr)NiO ₄ Nd ₂ NiO ₄ Ag-La(Sr)MnO ₃	La(Sr)Cr(Mn)O ₃ La(Sr)Cr(Ru)O ₃ Sr(La)TiO ₃ CuYSZ(CeO ₂)	Kerosene, Gasoline, Diesel, Alcohol, Biogas/ biofuel, Coal gas, DME	Automotive APU/power units, Stational applications with kerosene, Coal gas power generation, Portable applications, Direct-alcohol SOFCs, Carbon-neutral power generation using biofuels

have potential for wider varieties of applications in the future. Among others, the following possibilities should be emphasized to consider future directions: (1) High electrical efficiency beyond 50% is attractive to bring about distributed power generation [291], where high ionic conductivity of the electrolyte, low electrode polarization, low contact resistance, as well as sufficient durability are essential. (2) Low-temperature SOFCs may actualize mobile applications of SOFCs where much higher ionic conductance and electrode activities should be tailored [15, 292–297] and more sophisticated materials design at microscopic levels becomes important [9, 13, 298–300]. (3) As the operational temperature of the low-temperature SOFC approaches that of high-temperature PEFCs, we may also learn much from materials and electrochemical research for PEFCs [301–305], which may lead to an integration of materials research between SOFC and PEFC. (4) Fuel-flexible SOFCs may contribute to energy security by using a wide variety of fuels and carbon-neutral power generation by using regenerable energy resources, for which robustness and poisoning tolerance of electrodes may be critical [255, 260, 262]. SOFC researchers should take responsibilities to fulfill their potentials and possibilities via concept demonstrations of operating materials, devices, and systems.

Readers may read references 9, 21, 23, 24, 26, 28, 29, 79, and 100 to understand fundamental materials science related to the topic of this chapter. In order to understand the very early stage of this research field, readers may check references 39, 40, and 126.

Acknowledgments

Financial support by the current NEDO project “Development of System and Elemental Technology on Solid Oxide Fuel Cells (SOFC)” is gratefully acknowledged. We thank Associate Prof. A. Hayashi and WPI Assistant Profs. S. Lyth and S. Bishop in Kyushu University for carefully reading the manuscript and Ms. A. Zaitzu in Kyushu University for her editorial work.

Bibliography

- Steele BCH, Heinzel A (2001) Materials for fuel-cell technologies. *Nature* 414:345–352
- Minh NQ, Takahashi T (1995) Science and technology of ceramic fuel cells. Elsevier, Amsterdam
- Singhal SC, Kendall K (2003) High-temperature solid oxide fuel cells: fundamentals, design and application. Elsevier, Oxford, UK
- Larminie J, Dicks A (2003) Fuel cell systems explained, 2nd edn. Wiley, West Sussex, UK
- Tagawa H (1998) Solid oxide fuel cells and global environment. Agune-Shofudo, Shinjuku, Tokyo (in Japanese)
- Steinberger-Wilckens R (2009) European SOFC R&D – status and trends. *ECS Trans* 25(2):3–10
- Hosoi K, Nakabaru M (2009) Status of national project for SOFC development in Japan. *ECS Trans* 25(2):11–20
- Surdovall WA (2009) The status of SOFC programs in USA 2009. *ECS Trans* 25(2):21–27
- Maier J (2004) Physical Chemistry of ionic materials: ions and electrons in solids. Wiley, West Sussex, UK
- Sata N, Eberman K, Eberl K, Maier J (2000) Mesoscopic fast ion conduction in nanometre-scale planar heterostructures. *Nature* 408:946–949
- Garcia-Barriocanal J, Rivera-Calzada A, Varela M, Sefrioui Z, Iborra E, Leon C, Pennycook SJ, Santamaria J (2008) Colossal ionic conductivity at interfaces of epitaxial $\text{ZrO}_2\text{:Y}_2\text{O}_3/\text{SrTiO}_3$ heterostructures. *Science* 321(5889):676–680
- Tuller HL (2000) Ionic conduction in nanocrystalline materials. *Solid State Ionics* 131(1–2):143–157
- Maier J (2004) Nano-ionics: more than just a fashionable slogan. *J Electroceram* 13(1–3):593–598
- Araki W, Arai Y (2010) Oxygen diffusion in yttria-stabilized zirconia subjected to uniaxial stress. *Solid State Ionics* 181(8–10):441–446
- Litzelman SJ, Hertz JL, Jung W, Tuller HL (2008) Opportunities and challenges in materials development for thin film solid oxide fuel cells. *Fuel Cells* 8(5):294–302
- Evans A, Bieberle-Hütter A, Rupp JLM, Gauckler LJ (2009) Review on microfabricated micro-solid oxide fuel cell membranes. *J Power Sources* 194(1):119–129
- Ihara M, Hasegawa S (2006) Quickly rechargeable direct carbon solid oxide fuel cell with propane for recharging. *J Electrochem Soc* 153(8):A1544–A1546
- Hauch A, Jensen SH, Ramousse S, Mogensen M (2006) Performance and durability of solid oxide electrolysis cells. *J Electrochem Soc* 153(9):A1741–A1747
- Ishihara T, Kanno T (2010) Steam electrolysis using LaGaO_3 based perovskite electrolyte for recovery of unused heat energy. *ISIJ Int* 50:1291–1295
- Minh NQ (2010) Operating characteristics of solid oxide fuel cell stacks and systems. *ECS Trans* 25(2):241–246
- Steele BCH (1989) Oxygen ion conductors. In: Takahashi T (ed) High conductivity solid ionic conductors. World Scientific, Singapore, pp 402–446
- Steele BCH (1992) Oxygen ion conductors and their technological applications. In: Balkanski M, Takahashi T, Tuller HL (eds) Solid state ionics. Elsevier, Amsterdam, pp 17–28
- Nowick AS (1984) Atom transport in oxides of the fluorite structure. In: Diffusion in crystalline solids. Academic, Orlando, pp 143–188

24. Hayes W, Stoneham AM (1985) Defects and defect processes in nonmetallic solids. Wiley, New York
25. Kilner JA, Steele BCH (1981) Mass transport in anion-deficient fluorite oxides. In: Sørensen OT (ed) Nonstoichiometric oxides. Academic, New York, pp 233–269
26. Kofstad P (1972) Nonstoichiometry, diffusion, and electrical conductivity in binary metal oxides. Wiley-Interscience, New York
27. Eyring L (1979) The binary rare earth oxides. In: Gschneidner KA Jr, Eyring L (eds) Handbook on the physics and chemistry of rare earths, vol 3. North-Holland, Amsterdam, pp 337–399, Chap 27
28. Tuller HL (1991) Highly conductive ceramics. In: Buchanan RC (ed) Ceramic materials for electronics. Marcel Dekker, New York, pp 379–433
29. Tuller HL (1981) Mixed conduction in nonstoichiometric oxides. In: Sørensen OT, Sørensen OT (eds) Nonstoichiometric oxides. Academic, New York, pp 271–335
30. Shimizu K, Kashiwagi K, Nishiyama H, Kakimoto S, Sugaya S, Yokoi H, Satsuma A (2008) Impedancemetric gas sensor based on Pt and WO₃ co-loaded TiO₂ and ZrO₂ as total NO_x sensing materials. Sens Actuators B Chem 130:707–712
31. Gong J, Li Y, Tang Z, Zhang Z (2000) Enhancement of the ionic conductivity of mixed calcia/yttria stabilized zirconia. Mater Lett 46:115–119
32. Gauckler LJ, Sasaki K (1994) Ionic and electronic conductivities of homogeneous and heterogeneous materials in the system ZrO₂-In₂O₃. Solid State Ionics 75:203–210
33. Sasaki K (1993) Phase equilibria, electrical conductivity, and electrochemical properties of ZrO₂-In₂O₃. Dissertation, Swiss Federal Institute of Technology, Zürich, Switzerland
34. Stafford RJ, Rothman SJ, Routbort JL (1989) Effect of dopant size on the ionic conductivity of cubic stabilized ZrO₂. Solid State Ionics 37:67–72
35. Arachi Y, Sakai H, Yamamoto O, Takeda Y, Imanishi N (1999) Electrical conductivity of the ZrO₂-Ln₂O₃ (Ln = lanthanides) system. Solid State Ionics 121:133–139
36. Patterson JW (1971) Conduction domains for solid electrolytes. J Electrochem Soc 118:1033–1039
37. Heyne L, Engleson D (1977) The speed of response of solid electrolyte galvanic cells for Gas sensing. J Electrochem Soc 124:727–735
38. Tuller HL, Moon PK (1988) Fast ion conductors: future trends. Mater Sci Eng B1:171–191
39. Nernst W (1899) Über die elektrolytische Leitung fester Körper bei sehr hohen Temperaturen. Z Elektrochem 6(2):41–43
40. Baur E, Preis H (1937) Über Brennstoff-Ketten mit Festleitern. Z Elektrochem 44(9):695–698
41. Sasaki K, Maier J (2000) Re-analysis of defect equilibria and transport properties of Y₂O₃-stabilized ZrO₂ using EPR and optical relaxation. Solid State Ionics 134(3–4):303–321
42. Inaba H, Tagawa H (1996) Ceria-based solid electrolytes. Solid State Ionics 83:1–16
43. Sammes NM, Tompsett GA, Näfe H, Aldinger F (1999) Bismuth based oxide electrolytes-structure and ionic conductivity. J Eur Ceram Soc 19:1801–1826
44. Mogensen M, Sammes NM, Tompsett GA (2000) Physical, chemical and electrochemical properties of pure and doped ceria. Solid State Ionics 129:63–94
45. Yamamoto O, Arachi Y, Sakai H, Takeda Y, Imanishi N, Mizutani Y, Kawai M, Nakamura Y (1998) Zirconia based oxide ion conductors for solid oxide fuel cells. Ionics 4:403–408
46. Arachi Y, Asai T, Yamamoto O, Takeda Y, Imanishi N, Kawate K, Tamakoshi C (2001) Electrical conductivity of ZrO₂-Sc₂O₃ doped with HfO₂, CeO₂, and Ga₂O₃. J Electrochem Soc 148(5):A520–A523
47. Yamamoto O, Arati Y, Takeda Y, Imanishi N, Mizutani Y, Kawai M, Nakamura Y (1995) Electrical conductivity of stabilized zirconia with ytterbia and scandia. Solid State Ionics 79:137–142
48. Tietz F, Fischer W, Hauber Th, Mariotto G (1997) Structural evolution of Sc-containing zirconia electrolytes. Solid State Ionics 100:289–295
49. Wang Z, Cheng M, Bi Z, Dong Y, Zhang H, Zhang J, Feng Z, Li C (2005) Structure and impedance of ZrO₂ doped with Sc₂O₃ and CeO₂. Mater Lett 59:2579–2582
50. Gödickemeier M, Michel B, Orliukas A, Bohac P, Sasaki K, Gauckler LJ, Heinrich H, Schwander P, Kistorz G, Hofmann H, Frei O (1994) Effect of intergranular glass films on the electrical conductivity of 3Y-TZP. J Mater Res 9(5):1228–1240
51. Biebler-Hütter A, Beckel D, Infortuna A, Muecke UP, Rupp JLM, Gauckler LJ, Rey-Mermet S, Murali P, Bieri NR, Hotz N, Stutz MJ, Poulikakos D, Heeb P, Müller P, Bernard A, Gmür R, Hocker T (2008) A micro-solid oxide fuel cell system as battery replacement. J Power Sources 177:123–130
52. Ji Y, Kilner JA, Carolan MF (2005) Electrical properties and oxygen diffusion in yttria-stabilised zirconia (YSZ)-La_{0.8}Sr_{0.2}MnO_{3±δ}(LSM) composites. Solid State Ionics 176:937–943
53. Ralph JM, Rossignol C, Kumar R (2003) Cathode materials for reduced-temperature SOFCs. J Electrochem Soc 150(11):A1518–A1522
54. Stochniol G, Syskakis E, Naoumidis A (1995) Chemical compatibility between strontium-doped lanthanum manganite and yttria-stabilized zirconia. J Am Ceram Soc 78:929–932
55. Zhao H, Huo I, Sun L, Yu L, Gao S, Zhao J (2004) Preparation, chemical stability and electrochemical properties of LSCF-CBO composite cathodes. Mater Chem Phys 88:160–166
56. Yahiro H, Eguchi K, Arai H (1986) Ionic conduction and micro-structure of the ceria-strontia system. Solid State Ionics 21:37–47
57. Thangadurai V, Kopp P (2007) Chemical synthesis of Ca-doped CeO₂ - intermediate temperature oxide ion electrolytes. J Power Sources 168:178–183
58. Eguchi K, Setoguchi T, Inoue T, Arai H (1992) Electrical properties of ceria-based oxides and their application to solid oxide fuel cells. Solid State Ionics 52:165–172
59. Yahiro H, Baba Y, Eguchi Y, Eguchi K, Arai H (1988) High temperature fuel cell with ceria-yttria solid electrolyte. J Electrochem Soc 135:2077–2080
60. Shobit Omar S, Wachsmann ED, Nino JC (2008) Higher conductivity Sm³⁺ and Nd³⁺ co-doped ceria-based electrolyte materials. Solid State Ionics 178:1890–1897

61. Bance P, Brandon NP, Girvan B, Holbeche P, O'Dea S, Steele BCH (2004) Spinning-out a fuel cell company from a UK University-2 years of progress at Ceres Power. *J Power Sources* 131:86–90
62. Göedickemeier M, Gauckler LJ (1998) Engineering of solid oxide fuel cells with ceria-based electrolytes. *J Electrochem Soc* 145:414–421
63. Atkinson A (1997) Chemically-induced stresses in gadolinium-doped ceria solid oxide fuel cell electrolytes. *Solid State Ionics* 95:249–258
64. Marques FMB, Navarro LM (1997) Performance of double layer electrolyte cells Part II: GCO/YSZ, a case study. *Solid State Ionics* 100:29–38
65. Tsoda A, Gupta A, Naoumoudis A, Skarmoutsos D, Nikolopoulos P (1998) Performance of a double-layer CGO/YSZ electrolyte for solid oxide fuel cells. *Ionics* 4:234–240
66. Kharton VV, Figueiredo FM, Navarro L, Naumovich EN, Kovalevsky AV, Yaremchenko AA, Viskup AP, Carneiro A, Marques FMB, Frade JR (2001) Ceria-based materials for solid oxide fuel cells. *J Mater Sci* 36:1105–1117
67. Oishi N, Atkinson A, Brandon NP, Kilner JA, Steele BCH (2005) Fabrication of an anode-supported gadolinium-doped ceria solid oxide fuel cell and its operation at 550°C. *J Am Ceram Soc* 88(6):1394–1396
68. Oishi N, Yoo Y (2010) Evaluation of metal supported ceria based solid oxide fuel cell fabricated by wet powder spray and sintering. *J Electrochem Soc* 157(1):B125–B129
69. Moon PK, Tuller HL (1990) Evaluation of the $Gd_2(Zr_xTi_{1-x})_2O_7$ pyrochlore system as an oxygen Gas sensor. *Sens Actuators B Chem* 1:199–202
70. Yamamura H, Nishino H, Kakinuma K, Nomura K (2003) Electrical conductivity anomaly around fluorite–pyrochlore phase boundary. *Solid State Ionics* 158:359–365
71. Moon PK, Tuller HL (1988) Ionic-conduction in the $Gd_2Ti_2O_7$ - $Gd_2Zr_2O_7$ system. *Solid State Ionics* 28–30(1):470–474
72. Kramer SA, Tuller HL (1995) A novel titanate-based oxygen ion conductor: $Gd_2Ti_2O_7$. *Solid State Ionics* 82:15–23
73. Kharton VV, Tsipis EV, Yaremchenko AA, Vyshatko NP, Shaula AL, Naumovich EN, Frade JR (2003) Oxygen ionic and electronic transport in $Gd_{2-x}Ca_xTi_2O_{7-8}$ pyrochlores. *J Solid State Electrochem* 7:468–476
74. Mori M, Tompsett GM, Sammes NM, Suda E, Takeda Y (2003) Compatibility of $Gd_xTi_2O_7$ pyrochlores ($1.72 \leq x \leq 2.0$) as electrolytes in high-temperature solid oxide fuel cells. *Solid State Ionics* 158:79–90
75. Yasuda I, Matsuzaki Y, Yamakawa T, Koyama T (2000) Electrical conductivity and mechanical properties of alumina-dispersed doped lanthanum gallates. *Solid State Ionics* 135:381–388
76. Tietz F (1999) Thermal expansion of SOFC materials. *Ionics* 5:129–139
77. Lee JH, Yoshimura M (1999) Phase stability and electrical conductivity of the $Zr_{0.5}Y_{0.5}O_{1.75}$ - $Y_{0.75}Nb_{0.25}O_{1.75}$ system. *Solid State Ionics* 124:185–191
78. Lee JH, Yashima M, Yoshimura M (1998) Ionic conductivity of fluorite-structured solid solution $Y_{0.8}Nb_{0.2}O_{1.7}$. *Solid State Ionics* 107:47–51
79. West AR (1999) Basic solid state chemistry. Wiley, Chichester
80. Roth RS (1957) Classification of perovskite and other ABO_3 -type compounds. *J Res Nat Bur Stand* 58:75–88
81. Takahashi T, Iwahara H (1971) Ionic conduction in perovskite-type oxide solid solution and its application to the solid electrolyte fuel cell. *Energy Convers* 11:105–111
82. Ishihara T, Matsuda H, Takita Y (1994) Doped $LaGaO_3$ perovskite type oxide as a new oxide ionic conductor. *J Am Chem Soc* 116:3801–3803
83. Ishihara T, Matsuda H, Takita Y (1995) Effects of rare earth cations doped for La site on the oxide ionic conductivity of $LaGaO_3$ -based perovskite type oxide. *Solid State Ionics* 79:147–151
84. Ishihara T, Ishikawa S, Yu C, Akbay T, Hosoi K, Nishiguchi H, Takita Y (2003) Oxide ion and electronic conductivity in Co doped $La_{0.8}Sr_{0.2}Ga_{0.8}Mg_{0.2}O_3$ perovskite oxide. *Phys Chem Chem Phys* 5:2257–2263
85. Ishihara T (2001) Current status of intermediate temperature solid oxide fuel cell (in Japanese). *Bull Ceram Soc Jpn* 36:483–485
86. Islam MS, Davis RA (2004) Atomistic study of dopant site-selectivity and defect association in the lanthanum gallate perovskite. *J Mater Chem* 14:86–93
87. Huang K, Feng M, Goodenough JB, Schmerling M (1996) Characterization of Sr-doped $LaMnO_3$ and $LaCoO_3$ as cathode materials for a doped $LaGaO_3$ ceramic fuel cell. *J Electrochem Soc* 143(11):3630–3636
88. Kostoglou GC, Ftikos C, Ahmad-Khanlou A, Naoumidis A, Stöver D (2000) Chemical compatibility of alternative perovskite oxide SOFC cathodes with doped lanthanum gallate solid electrolyte. *Solid State Ionics* 134:127–138
89. Shaula AL, Kharton VV, Marques FMB (2004) Phase interaction and oxygen transport in $La_{0.8}Sr_{0.2}Fe_{0.8}Co_{0.2}O_{3-x}(La_{0.9}Sr_{0.1})_{0.98}Ga_{0.8}Mg_{0.2}O_3$ composites. *J Eur Ceram Soc* 24:2631–2639
90. Sakai N, Horita T, Yamaji K, Brito ME, Yokokawa H, Kawakami A, Matsuoka S, Watanabe N, Ueno A (2006) Interface stability among solid oxide fuel cell materials with perovskite structures. *J Electrochem Soc* 153(3):A621–A625
91. Joshi AV, Steppan JJ, Taylor DM, Elangovan S (2004) Solid electrolyte materials, devices, and applications. *J Electroceram* 13:619–625
92. Schober T (1998) Protonic conduction in $BaIn_{0.5}Sn_{0.5}O_{2.75}$. *Solid State Ionics* 109:1–11
93. Goodenough JB (1997) Ceramic solid electrolytes. *Solid State Ionics* 94:17–25
94. Goodenough JB, Ruiz-Dias JE, Zhen YS (1990) Oxide-ion conduction in $Ba_2In_2O_5$ and $Ba_3In_2MO_8$ ($M = Ce, Hf, \text{ or } Zr$). *Solid State Ionics* 44:21–31
95. Manthiram A, Kuo JF, Goodenough JB (1993) Characterization of oxygen-deficient perovskites as oxide-ion electrolytes. *Solid State Ionics* 62:225–234

96. Uchimoto Y, Kinuhata M, Takagi H, Yao T, Inagaki T, Yoshida H (1999) Crystal structure of metal cation-doped $\text{Ba}_2\text{In}_2\text{O}_5$ and its oxide ion conductivity. In: Proceedings of SOFC VI, Electrochemical Society, Pennington, pp 317–326
97. Schober T, Friedrich J, Krug F (1997) Phase transition in the oxygen and proton conductor $\text{Ba}_2\text{In}_2\text{O}_5$ in humid atmospheres below 300°C. *Solid State Ionics* 99:9–13
98. Hashimoto T, Inagaki Y, Kishi A, Dokiya M (2000) Absorption and secession of H_2O and CO_2 on $\text{Ba}_2\text{In}_2\text{O}_5$ and their effects on crystal structure. *Solid State Ionics* 128:227–231
99. Zhang GB, Smyth DM (1995) Defects and transport of the brownmillerite oxides with high oxygen ion conductivity - $\text{Ba}_2\text{In}_2\text{O}_5$. *Solid State Ionics* 82:161–172
100. Kingery WD, Bowen HK, Uhlmann DR (1976) Introduction to ceramics, 2nd edn. Wiley, New York
101. Oyane A (2010) Development of apatite-based composites by a biomimetic process for biomedical applications. *J Ceram Soc Jpn* 118(2):77–81
102. Felsche J (1972) Rare earth silicates with the apatite structure. *J Solid State Chem* 5:266–275
103. Park J, Lakes RS (2007) Biomaterials: an introduction, 3rd edn. Springer, New York
104. Bonder IA (1982) Rare-earth silicates. *Ceramics Int* 8(3):83–89
105. Higuchi Y, Sugawara M, Onishi K, Sakamoto M, Nakayama S (2010) Oxide ionic conductivities of apatite-type lanthanum silicates and germanates and their possibilities as an electrolyte of lower temperature operating SOFC. *Ceramics Int* 36:955–959
106. Panteix PJ, Béchade E, Julien I, Abélard P, Bernache-Assollant D (2008) Influence of anionic vacancies on the ionic conductivity of silicated rare earth apatites. *Mater Res Bull* 43: 1223–1231
107. Yoshioka H (2006) Oxide ionic conductivity of apatite-type lanthanum silicates. *J Alloys Comp* 408–412:649–652
108. Higuchi M, Masubuchi Y, Nakayama S, Kikkawa S, Kodaira K (2004) Single crystal growth and oxide ion conductivity of apatite-type rare-earth silicates. *Solid State Ionics* 174:73–80
109. Masubuchi Y, Higuchi M, Takeda T, Kikkawa S (2006) Preparation of apatite-type $\text{La}_{9.33}(\text{SiO}_4)_6\text{O}_2$ oxide ion conductor by alcoxide-hydrolysis. *J Alloys Comp* 408–412:641–644
110. Nojiri Y, Tanase S, Iwasa M, Yoshioka H, Matsumura Y, Sakai T (2010) Ionic conductivity of apatite-type solid electrolyte material, $\text{La}_{10-x}\text{Ba}_x\text{Si}_6\text{O}_{27-x/2}$ ($x = 0-1$), and its fuel cell performance. *J Power Sources* 195:4059–4064
111. Mineshige A, Nakao T, Kobune M, Yazawa T, Yoshioka H (2008) Electrical properties of $\text{La}_{10}\text{Si}_6\text{O}_{27}$ -based oxides. *Solid State Ionics* 179:1009–1012
112. Nakao T, Mineshige A, Kobune M, Yazawa T, Yoshioka H (2008) Chemical stability of $\text{La}_{10}\text{Si}_6\text{O}_{27}$ and its application to electrolytes for solid oxide fuel cells. *Solid State Ionics* 179:1567–1569
113. Nakayama S, Kageyama T, Aono H, Sadaoka Y (1995) Ionic conductivity of lanthanoid silicates, $\text{Ln}_{10}(\text{SiO}_4)_6\text{O}_3$ ($\text{Ln} = \text{La}, \text{Nd}, \text{Sm}, \text{Gd}, \text{Dy}, \text{Y}, \text{Ho}, \text{Er}$ and Yb). *J Mater Chem* 5:1801–1805
114. Nakayama S, Sakamoto M (1998) Electrical properties of new type high oxide ionic conductor $\text{RE}_{10}\text{Si}_6\text{O}_{27}$ ($\text{RE} = \text{La}, \text{Pr}, \text{Nd}, \text{Sm}, \text{Gd}, \text{Dy}$). *J Euro Ceram Soc* 18:1413–1418
115. Yoshioka H (2007) Enhancement of ionic conductivity of apatite-type lanthanum silicates doped with cations. *J Am Ceram Soc* 90:3099–3105
116. Yoshioka H, Nojiri Y, Tanase S (2008) Ionic conductivity and fuel cell properties of apatite-type lanthanum silicates doped with Mg and containing excess oxide ions. *Solid State Ionics* 179:2165–2169
117. Li B, Liu W, Pan W (2010) Synthesis and electrical properties of apatite-type $\text{La}_{10}\text{Si}_6\text{O}_{27}$. *J Power Sources* 195:2196–2201
118. Pivak YV, Kharton VV, Yaremchenko AA, Yakovlev SO, Kovalevsky AV, Frade JR, Marques FMB (2007) Phase relationships and transport in Ti-, Ce- and Zr-substituted lanthanum silicate systems. *J Eur Ceram Soc* 27:2445–2454
119. Nojiri Y, Chen WF, Tanase S, Iwasa M, Matsumura Y, Sakai T, Tanase S (2008) Lanthanum silicate with apatite-type structure as an electrolyte for intermediate temperature SOFCs and the electrode materials. *ITE-IBA Lett* 1(6):498–506
120. León-Reina L, Porras-Vázquez JM, Losilla ER, Aranda MAG (2007) Phase transition and mixed oxide-proton conductivity in germanium oxy-apatites. *J Solid State Chem* 180:1250–1258
121. Arikawa H, Nishiguchi H, Ishihara T, Takita Y (2000) Oxide ion conductivity in Sr-doped $\text{La}_{10}\text{Ge}_6\text{O}_{27}$ apatite oxide. *Solid State Ionics* 136–137:31–37
122. Ishihara T, Arikawa H, Akbay T, Nishiguchi H, Takita Y (2001) Nonstoichiometric $\text{La}_{2-x}\text{GeO}_{5-\delta}$ monoclinic oxide as a new fast oxide ion conductor. *J Am Chem Soc* 123:203–209
123. Nakayama S, Higuchi Y, Kondo Y, Sakamoto M (2004) Effects of cation- or oxide ion-defect on conductivities of apatite-type La-Ge-O system ceramics. *Solid State Ionics* 170: 219–223
124. Berastegui P, Hull S, García García FJ, Grins J (2002) A structural investigation of $\text{La}_2(\text{GeO}_4)\text{O}$ and alkaline-earth-doped $\text{La}_{9.33}(\text{GeO}_4)_6\text{O}_2$. *J. Solid State Chem* 168:294–305
125. Kreuer KD (1996) Proton conductivity: materials and applications. *Chem Mater* 8(3):610–641
126. Wagner C (1968) Die Löslichkeit von Wasserdampf in $\text{ZrO}_2\text{-Y}_2\text{O}_3$ -Mischkristallen, *Berichte Bunseng. Phys Chem* 72(7):778–781
127. Alberti G, Casciola M (2001) Solid state protonic conductors, present main applications and future prospects. *Solid State Ionics* 145:3–16
128. Xu ZP, Jin Y, Diniz da Costa JC, Lu GQM (2008) $\text{Zr}(\text{HPO}_4)_2$ based organic/inorganic nanohybrids as new proton conductors. *Solid State Ionics* 178:1654–1659
129. Ahmad MI, Zaidi SMJ, Ruhman SU, Ahmed S (2006) Synthesis and proton conductivity of heteropolyacids loaded Y-zeolite as solid proton conductors for fuel cell applications. *Micropor Mesopor Mater* 91:296–304
130. Norby T (1999) Solid-state protonic conductors: principles, properties, progress and prospects. *Solid State Ionics* 125:1–11
131. Daiko Y, Nguyen VH, Yazawa T, Muto H, Sakai M, Matsuda A (2010) Phase transition and proton conductivity of

- CsHSO₄-WPA composites prepared by mechanical milling. *Solid State Ionics* 181:183–186
132. Sugahara T, Hayashi A, Tadanaga K, Tatsumisago M (2010) Characterization of proton conducting CsHSO₄-CsH₂PO₄ ionic glasses prepared by the melt-quenching method. *Solid State Ionics* 181:190–192
 133. Iwahara H, Esaka T, Uchida H, Maeda N (1981) Proton conduction in sintered oxides and its application to steam electrolysis for hydrogen production. *Solid State Ionics* 3(4):359–363
 134. Fukatsu N, Kurita N, Yajima T, Koide K, Ohashi T (1995) Proton conductors of oxide and their application to research into metal-hydrogen systems. *J Alloys Comp* 231:706–712
 135. Schwartz M, Link BF, Sammells AF (1993) New brownmillerite solid electrolytes. *J Electrochem Soc* 140:L62–L63
 136. Zahang GB, Smyth DM (1995) Protonic conduction in Ba₂In₂O₅. *Solid State Ionics* 82:153–160
 137. Fisher CSJ, Islam MS (1999) Defect, protons and conductivity in brownmillerite-structured Ba₂In₂O₅. *Solid State Ionics* 118:355–363
 138. Orera A, Slater PR (2010) Water incorporation studies in apatite-type rare earth silicates/germinates. *Solid State Ionics* 181:110–114
 139. Marrero-López D, Martín-Sedeño MC, Ruiz-Morales JC, Núñez P, Ramos-Barrado JR (2010) Preparation and characterisation of La_{10-x}Ge_{5.5}Al_{0.5}O_{26±δ} apatites by freeze-drying precursor method. *Mater Res Bull* 45:409–415
 140. Ito N, Iijima M, Kimura K, Iguchi S (2005) New intermediate temperature fuel cell with ultra-thin proton conductor electrolyte. *J Power Sources* 152:200–203
 141. Matsumoto H, Nomura I, Okada S, Ishihara T (2008) Intermediate-temperature solid oxide fuel cells using perovskite-type oxide based on barium cerate. *Solid State Ionics* 179:1486–1489
 142. Ishigaki T, Yamauchi S, Kishio K, Fueki K, Iwahara H (1986) Dissolution of deuterium into proton conductor SrCe_{0.95}Yb_{0.05}O_{3-δ}. *Solid State Ionics* 21:239–241
 143. Iwahara H, Uchida H, Ono K, Ogaki K (1998) Proton conduction in sintered oxides based on BaCeO₃. *J Electrochem Soc* 135:529–533
 144. Bonanos N, Ellis B, Mahmood MN (1991) Construction and operation of fuel cells based on the solid electrolyte BaCeO₃: Gd. *Solid State Ionics* 44:305–311
 145. Ma G, Matsumoto H, Iwahara H (1999) Ionic conduction and nonstoichiometry in non-doped Ba_xCeO_{3-α}. *Solid State Ionics* 122:237–247
 146. Taniguchi N, Hatoh K, Niikura J, Gamo T, Iwahara H (1992) Proton conductive properties of gadolinium-doped barium cerates at high temperatures. *Solid State Ionics* 53–56:998–1003
 147. Ma G, Shimura T, Iwahara H (1998) Ionic conduction and nonstoichiometry in Ba_xCe_{0.90}Y_{0.10}O_{3-α}. *Solid State Ionics* 110:103–110
 148. Ma G, Shimura T, Iwahara H (1999) Simultaneous doping with La³⁺ and Y³⁺ for Ba²⁺- and Ce⁴⁺-sites in BaCeO₃ and the ionic conduction. *Solid State Ionics* 120:51–60
 149. Katahira K, Kohchi Y, Shimura T, Iwahara H (2000) Protonic conduction in Zr-substituted BaCeO₃. *Solid State Ionics* 138:91–98
 150. Ranran P, Yan W, Lizhai Y, Zongqiang M (2006) Electrochemical properties of intermediate-temperature SOFCs based on proton conducting Sm-doped BaCeO₃ electrolyte thin film. *Solid State Ionics* 177:389–393
 151. Su XT, Yan QZ, Ma YH, Zhang WF, Ge CC (2006) Effect of co-dopant addition on the properties of yttrium and neodymium doped barium cerate electrolyte. *Solid State Ionics* 177:1041–1045
 152. Gorbova E, Zhuravlev BV, Demin AK, Song SQ, Tsiakaras PE (2006) Charge transfer properties of BaCe_{0.88}Nd_{0.12}O_{3-δ} co-ionic electrolyte. *J Power Sources* 157:720–723
 153. Taherparvar H, Kilner JA, Baker RT, Sahibzada M (2003) Effect of humidification at anode and cathode in proton-conducting SOFCs. *Solid State Ionics* 162–163:297–303
 154. Du Y, Nowick AS (1995) Structural transitions and proton conduction in nonstoichiometric A₃B'B₂O₉ perovskite-type oxides. *J Am Ceram Soc* 78:3033–3039
 155. Murugaraj P, Kreuer KD, He T, Schober T, Maier J (1997) High proton conductivity in barium yttrium stannate Ba₂YSnO_{5.5}. *Solid State Ionics* 98:1–6
 156. Kreuer KD (2003) Proton conducting oxides. *Ann Rev Mater Res* 33:333–359
 157. He T, Kreuer KD, Baikov YM, Maier J (1997) Impedance spectroscopic study of thermodynamics and kinetics of Gd-doped BaCeO₃ single crystal. *Solid State Ionics* 95:301–308
 158. Fabbri E, D'Epifanio A, Di Bartolomeo E, Licoccia S, Traversa E (2008) Tailoring the chemical stability of Ba(Ce_{0.8-x}Zr_x)Y_{0.2}O_{3-δ} protonic conductors for intermediate temperature solid oxide fuel cells (IT-SOFCs). *Solid State Ionics* 179:558–564
 159. Kreuer KD (1999) Aspects of the formation and mobility of protonic charge carriers and the stability of perovskite-type oxides. *Solid State Ionics* 125:285–302
 160. He T, Ehrhart P, Meuffels P (1995) Optical band gap and Urbach tail in Y-doped BaCeO₃. *J Appl Phys* 79:3219–3223
 161. Minh NQ (1993) Ceramic fuel cells. *J Am Ceram Soc* 76(3):563–588
 162. Mackor A, Koster TPM, Kraaijkamp JG, Gerretsen J, van Eijk JPGM (1991) Influence of La-substitution and -substoichiometry on conductivity, thermal expansion and chemical stability of Ca- or Sr-doped lanthanum manganites as SOFC cathodes. In: Grosz F, Zegers P, Singhal SC, Yamamoto O (eds) *Proceedings of the 2nd international symposium of solid oxide fuel cells*. Commission of European Communities, Luxembourg, pp 463–471
 163. Anderson HU (1992) Review of p-type doped perovskite materials for SOFC and other applications. *Solid State Ionics* 52:33–41
 164. Kuo JH, Anderson HU, Sparlin DM (1989) Oxidation-reduction behavior of undoped and Sr-doped LaMnO₃: nonstoichiometry and defect structure. *J Solid State Chem* 83:52–60
 165. Kuo JH, Anderson HU, Sparlin DM (1990) Oxidation-reduction behavior of undoped and Sr-doped LaMnO₃: defect

- structure, electrical conductivity, and thermoelectric power. *J Solid State Chem* 87:55–63
166. van Roosmalen JAM, Huijsmans JPP, Cordfunke EHP (1991) Sinter behavior and electrical conductivity of $(\text{La,Sr})\text{MnO}_3$ as a function of Sr-content. In: Grosz F, Zegers P, Singhal SC, Yamamoto O (eds) *Proceedings 2nd International symposium on solid oxide fuel cells*. Commission of European Communities, Luxembourg, pp 507–516
 167. Hammouche A, Siebert E, Hammou A (1989) Crystallographic, thermal and electrochemical properties of the system $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$ for high temperature solid electrolyte fuel cells. *Mater Res Bull* 24:367–380
 168. Yamada H, Nagamoto H (1993) Thermal expansion coefficient and electrical conductivity of Mn-based perovskite-type oxides. In: Singhal SC, Iwahara H (eds) *Proceedings of the 3rd International symposium on solid oxide fuel cells*, vol 93–94. The Electrochem Society Inc. Proceedings, Pennington, pp 213–219
 169. Nasrallah MM, Anderson HU, Stevenson JW (1991) Defect chemistry and properties of $\text{Y}_{1-x}\text{Ca}_x\text{MnO}_3$. In: Grosz F, Zegers P, Singhal SC, Yamamoto O (eds) *Proceedings of the 2nd international symposium on solid oxide fuel cells*. Commission of European Communities, Luxembourg, pp 545–552
 170. Stevenson JW, Nasrallah MM, Anderson HU, Parlin DMS (1993) Defect structure of $\text{Y}_{1-y}\text{Ca}_y\text{MnO}_3$ and $\text{La}_{1-y}\text{Ca}_y\text{MnO}_3$, I. Electrical properties. *J Solid State Chem* 102:175–184
 171. Stevenson JW, Nasrallah MM, Anderson HU, Sparlin DM (1993) Defect structure of $\text{Y}_{1-y}\text{Ca}_y\text{MnO}_3$ and $\text{La}_{1-y}\text{Ca}_y\text{MnO}_3$ II. Oxidation-reduction behavior. *J Solid State Chem* 102: 185–197
 172. Fu B, Huebner W, Trubelja MF, Stubican VS (1993) $(\text{Y}_{1-x}\text{Ca}_x)\text{FeO}_3$: a potential cathode material for solid oxide fuel cells. In: Singhal SC, Iwahara H (eds) *Proceedings of the 3rd international symposium on solid oxide fuel cells*. Proceedings, vol 93–4. The Electrochemical Society, Pennington, pp 276–287
 173. Yamamoto O, Takeda Y, Imanishi N, Sakaki Y (1993) Electrochemical properties of $\text{La}_{1-x}\text{Ca}_x\text{MnO}_{3-z}$ as cathode in SOFC. In: Singhal SC, Iwahara H (eds) *Proceedings of the 3rd international symposium on solid oxide fuel cells*. Proceedings, vol 93–4. The Electrochemical Society, Pennington, pp 205–212
 174. Mizusaki J, Tagawa H, Katou M, Hirano K, Sawata A, Tsuneyoshi K (1991) Electrochemical properties of some perovskite-type oxides as oxygen gas electrodes on yttria stabilized zirconia. In: Grosz F, Zegers P, Singhal SC, Yamamoto O (eds) *Proceedings of the 2nd international symposium on solid oxide fuel cells*. Commission of European Communities, Luxembourg, pp 487–494
 175. Mizusaki J, Tagawa H, Tsuneyoshi K, Sawata A (1991) Reaction kinetics and microstructure of the solid oxide fuel cells air electrode $\text{La}_{0.6}\text{Ca}_{0.4}\text{MnO}_3/\text{YSZ}$. *J Electrochem Soc* 138(7):1867–1873
 176. Dokiya M (1992) A historical review on the SOFC research activity at NCLII, (in Japanese). In: *Extended abstracts 1st symposium on solid oxide fuel cells*, Japan. Solid Oxide Fuel Cells Society Japan, Tokyo, pp 11–14
 177. Yokokawa H, Sakai N, Kawada T, Dokiya M (1991) Chemical thermodynamic compatibility of solid oxide fuel cell materials. In: Grosz F, Zegers P, Singhal SC, Yamamoto O (eds) *Proceedings of the 2nd international symposium on solid oxide fuel cells*. Commission of European Communities, Luxembourg, pp 663–670
 178. Yokokawa H, Sakai N, Kawada T, Dokiya M (1992) Thermodynamic stabilities of perovskite oxides for electrodes and other electrochemical materials. *Solid State Ionics* 52:43–56
 179. Otoshi S, Sasaki H, Ohnishi H, Hase M, Ishimaru K, Ippommatsu M, Higuchi T, Miyayama M, Yanagida H (1991) Changes in the phases and electrical conduction properties of $(\text{La}_{1-x}\text{Sr}_x)_{1-y}\text{MnO}_{3-\delta}$. *J Electrochem Soc* 138(5):1519–1523
 180. Takeda Y, Kanno R, Noda M, Tomida Y, Yamamoto O (1987) Cathodic polarization phenomena of perovskite oxide electrodes with stabilized zirconia. *J Electrochem Soc* 134(11):2656–2661
 181. Mizusaki J (1992) Nonstoichiometry, diffusion, and electrical properties of perovskite-type oxide electrode materials. *Solid State Ionics* 52:79–91
 182. Teraoka Y, Nobunaga T, Okamoto K, Miura N, Yamazoe N (1991) Influence of constituent metal cations in substituted LaCoO_3 on mixed conductivity and oxygen permeability. *Solid State Ionics* 48:207–212
 183. Ivers-Tiffée E, Schiefl M, Oel HJ, Wersing W (1993) Investigations of cobalt-containing perovskites in SOFC single cells with respect to interface reactions and cell performance. In: Singhal SC, Iwahara H (eds) *Proceedings of the 3rd international symposium on solid oxide fuel cells*. Proceedings, vol 93–94. The Electrochemical Society, Pennington, pp 613–622
 184. Iberl A, von Philipsborn H, Schiefl M, Ivers-Tiffée E, Wersing W, Zorn G (1991) High-temperature X-ray diffraction measurements of phase transitions and thermal expansion in $(\text{La,Sr})(\text{Mn,Co})\text{O}_3$ -cathode materials. In: Grosz F, Zegers P, Singhal SC, Yamamoto O (eds) *Proceedings of the 2nd international symposium on solid oxide fuel cells*. Commission of European Communities, Luxembourg, pp 527–535
 185. Mackor A, Spee CIMA, van der Zouwen-Assink EA, Baptista JL, Schoonman J (1990) Mixed conductivity in perovskite SOFC materials $\text{La}_{1-x}\text{M}_x\text{Mn}_{1-y}\text{Co}_y\text{O}_3$ ($\text{M} = \text{Ca}$ or Sr). In: *Proceedings of the 25th intersociety energy conversion engineering conference*, American Institute of Chemical Engineers, vol 3, Aug. pp 251–255
 186. Tai LW, Nasrallah MM, Anderson HU (1993) $(\text{La}_{1-x}\text{Sr}_x)(\text{Co}_{1-y}\text{Fe}_y)\text{O}_3$, A potential cathode for intermediate temperature SOFC applications. In: Singhal SC, Iwahara H (eds) *Proceedings of the 3rd international symposium on solid oxide fuel cells*. Proceedings, vol 93–4. The Electrochemical Society, Pennington, pp 241–251
 187. Chen CC, Nasrallah MM, Anderson HU (1993) Preparation and electrode characteristics of dense $\text{La}_{0.6}\text{Sr}_{0.4}\text{Co}_{0.2}\text{Fe}_{0.8}\text{O}_3$ thin film by polymeric precursors. In: Singhal SC, Iwahara H (eds)

- Proceedings of the 3rd international symposium on solid oxide fuel cells. Proceedings, vol 93–4. The Electrochemical Society Inc., Pennington, pp 252–266
188. Teraoka Y, Zhang HM, Okamoto K, Yamazoe N (1988) Mixed ionic-electronic conductivity of $\text{La}_{1-x}\text{Sr}_x\text{Co}_{1-y}\text{Fe}_y\text{O}_{3-\delta}$. *Mater Res Bull* 23:51–58
 189. Ftikos C, Carter S, Steele BCH (1993) Mixed electronic/ionic conductivity of the solid solutions $\text{La}_{(1-x)}\text{Sr}_x\text{Co}_{(1-y)}\text{Ni}_y\text{O}_{3-\delta}$ ($x:0.4, 0.5, 0.6$ and $y:0.2, 0.4, 0.6$). *J Europ Ceram Soc* 12:79–86
 190. Inoue T, Seki N, Eguchi K, Arai H (1990) Low-temperature operation of solid electrolyte oxygen sensors using perovskite-type oxide electrodes and cathodic reaction kinetics. *J Electrochem Soc* 137(8):2523–2527
 191. Sasaki K, Wurth JP, Gschwend R, Gödickemeier M, Gauckler LJ (1996) Microstructure-property relations of solid oxide fuel cell cathodes and current collectors: cathodic polarization and ohmic resistance. *J Electrochem Soc* 143:530–543
 192. Kleitz M (1992) Reaction pathways: a new electrode modelling concept. In: McEvoy A (ed) *Fundamental barriers to SOFC performance*. Swiss Federal Office of Energy, Bern, pp 4–12
 193. Steele BCH, Carter S, Kajda J, Kontoulis I, Kilner JA (1991) Optimisation of fuel cell components using $^{18}\text{O}/^{16}\text{O}$ exchange and dynamic SIMS techniques. In: Grosz F, Zegers P, Singhal SC, Yamamoto O (eds) *Proceedings of the 2nd international symposium on solid oxide fuel cells*. Commission of European Communities, Luxembourg, pp 517–525
 194. Hammou A (1992) Solid oxide fuel cells. In: Gerischer H, Tobias CW (eds) *Advances in electrochemical science and engineering*, vol 2. VCH-Verlag, Weinheim, pp 87–139
 195. Hammouche A, Siebert E, Hammou A, Kleitz M, Caneiro A (1991) Electrocatalytic properties and nonstoichiometry of the high temperature air electrode $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$. *J Electrochem Soc* 138(5):1212–1216
 196. Ishigaki T, Yamauchi S, Kishio K, Mizusaki J, Fueki K (1988) Diffusion of oxide ion vacancies in perovskite-type oxides. *J Solid State Chem* 73:179–187
 197. Takeda Y, Ueno H, Imanishi N, Yamamoto O, Sammes N, Phillipps M (1996) $\text{Gd}_{1-x}\text{Sr}_x\text{CoO}_3$ for the electrode of solid oxide fuel cells. *Solid State Ionics* 86–88:1187–1190
 198. Tu H, Takeda Y, Imanishi N, Yamamoto O (1997) $\text{Ln}_{1-x}\text{Sr}_x\text{CoO}_3$ ($\text{Ln} = \text{Sm}, \text{Dy}$) for the electrode of solid oxide fuel cells. *Solid State Ionics* 100:283–288
 199. Sakaki Y, Takeda Y, Kato A, Imanishi N, Yamamoto O, Hattori M, Iio M, Esaki Y (1999) $\text{Ln}_{1-x}\text{Sr}_x\text{MnO}_3$ ($\text{Ln} = \text{Pr}, \text{Nd}, \text{Sm}$ and Gd) as the cathode material for solid oxide fuel cells. *Solid State Ionics* 118:187–194
 200. Phillipps M, Sammes N, Yamamoto O (1999) $\text{Gd}_{1-x}\text{A}_x\text{Co}_{1-y}\text{Mn}_y\text{O}_3$ ($\text{A} = \text{Sr}, \text{Ca}$) as a cathode for the SOFC. *Solid State Ionics* 123:131–138
 201. Tu H, Takeda Y, Imanishi N, Yamamoto O (1999) $\text{Ln}_{0.4}\text{Sr}_{0.6}\text{Co}_{0.8}\text{Fe}_{0.2}\text{O}_{3-\delta}$ ($\text{Ln} = \text{La}, \text{Pr}, \text{Nd}, \text{Sm}, \text{Gd}$) for the electrode in solid oxide fuel cells. *Solid State Ionics* 117:277–281
 202. Qiu L, Ichikawa T, Hirano A, Imanishi N, Takeda Y (2003) $\text{Ln}_{1-x}\text{Sr}_x\text{Co}_{1-y}\text{Fe}_y\text{O}_{3-\delta}$ ($\text{Ln} = \text{Pr}, \text{Nd}, \text{Gd}; x = 0.2, 0.3$) for the electrodes of solid oxide fuel cells. *Solid State Ionics* 158:55–65
 203. Ishihara T, Honda M, Shibayama T, Nishiguchi H, Takita Y (1998) Intermediate temperature solid oxide fuel cells using a New LaGaO_3 based oxide ion conductor. *J Electrochem Soc* 145:3177–3183
 204. Ishihara T, Fukui S, Nishiguchi H, Takita Y (2002) Mixed electronic-oxide ionic conductor of BaCoO_3 doped with La for cathode of intermediate-temperature-operating solid oxide fuel cell. *Solid State Ionics* 152–153:609–613
 205. Shao ZP, Haile SM (2004) A high-performance cathode for the next generation of solid-oxide fuel cells. *Nature* 431:170–173
 206. Li S, Lu Z, Wei B, Huang X, Miao J, Cao G, Zhu R, Su W (2006) A study of $(\text{Ba}_{0.5}\text{Sr}_{0.5})_{1-x}\text{Sm}_x\text{Co}_{0.8}\text{Fe}_{0.2}\text{O}_{3-\delta}$ as a cathode material for IT-SOFCs. *J Alloys Comp* 426:408–414
 207. Kotomin EA, Mastrikov YA, Kuklja MM, Merkle R, Roytburd A, Maier J (2011) First principles calculations of oxygen vacancy formation and migration in mixed conducting $\text{Ba}_{0.5}\text{Sr}_{0.5}\text{Co}_{1-y}\text{Fe}_y\text{O}_{3-\delta}$ perovskites. *Solid State Ionics* 188:1–5
 208. Yan A, Cheng M, Dong Y, Yang W, Maragou V, Song S, Tsiakaras P (2006) Investigation of a $\text{Ba}_{0.5}\text{Sr}_{0.5}\text{Co}_{0.8}\text{Fe}_{0.2}\text{O}_{3-\delta}$ based cathode IT-SOFC I. The effect of CO_2 on the cell performance. *Appl Catal B Environ* 66:64–71
 209. Yan A, Yang M, Hou Z, Dong Y, Cheng M (2008) Investigation of $\text{Ba}_{1-x}\text{Sr}_x\text{Co}_{0.8}\text{Fe}_{0.2}\text{O}_{3-\delta}$ as cathodes for low-temperature solid oxide fuel cells both in the absence and presence of CO_2 . *J Power Sources* 185:76–84
 210. Taniguchi S, Kadowaki M, Kawamura H, Yasuo T, Akiyama Y, Miyake Y, Saitoh T (1995) Degradation phenomena in the cathode of a solid oxide fuel cell with an alloy separator. *J Power Sources* 55:73–79
 211. Taniguchi S, Kadowaki M, Yasuo T, Akiyama Y, Itoh Y, Miyake Y, Nishio K (1996) Suppression of chromium diffusion to an SOFC cathode from an alloy separator by a cathode second layer. *Denki Kagaku* 64(6):568–574
 212. Matsuzaki Y, Yasuda I (2001) Dependence of SOFC cathode degradation by chromium-containing alloy on compositions of electrodes and electrolytes. *J Electrochem Soc* 148(2):A126–A131
 213. Chiba R, Yoshimura F, Sakurai Y (1999) An investigation of $\text{LaNi}_{1-x}\text{Fe}_x\text{O}_3$ as a cathode material for solid oxide fuel cells. *Solid State Ionics* 124:281–288
 214. Komatsu T, Chiba R, Arai H, Sato K (2008) Chemical compatibility and electrochemical property of intermediate-temperature SOFC cathodes under Cr poisoning condition. *J Power Sources* 176:132–137
 215. Jiang SP, Zhen Y (2008) Mechanism of Cr deposition and its application in the development of Cr-tolerant cathodes of solid oxide fuel cells. *Solid State Ionics* 179:1459–1464
 216. Lee S, Bevilacqua M, Fornasiero P, Vohs JM, Gorte RJ (2009) Solid oxide fuel cell cathodes prepared by infiltration of $\text{LaNi}_{0.6}\text{Fe}_{0.4}\text{O}_3$ and $\text{La}_{0.91}\text{Sr}_{0.09}\text{Ni}_{0.6}\text{Fe}_{0.4}\text{O}_3$ in porous yttria-stabilized zirconia. *J Power Sources* 193:747–753

217. Skinner S, Kilner J (2000) Oxygen diffusion and surface exchange in $\text{La}_{2-x}\text{Sr}_x\text{NiO}_{4+\delta}$. *Solid State Ionics* 135:709–712
218. Mauvy F, Bassat J, Boehm E, Manaud J, Dordor P, Grenier J (2003) Oxygen electrode reaction on $\text{Nd}_2\text{NiO}_{4+\delta}$ cathode materials: impedance spectroscopy study. *Solid State Ionics* 158:17–28
219. Lalanne C, Prosperi G, Bassat J, Mauvy F, Fourcade S, Stevens P, Zahid M, Diethelm S, van Herle J, Grenier J (2008) Neodymium-deficient nickelate oxide $\text{Nd}_{1.95}\text{NiO}_{4+\delta}$ as cathode material for anode-supported intermediate temperature solid oxide fuel cells. *J Power Sources* 185:1218–1224
220. Nie H, Wen T, Wang S, Wang Y, Guth U, Vashook V (2006) Preparation, thermal expansion, chemical compatibility, electrical conductivity and polarization of $\text{A}_{2-\alpha}\text{A}'_\alpha\text{MO}_4$ ($\text{A} = \text{Pr}, \text{Sm}; \text{A}' = \text{Sr}; \text{M} = \text{Mn}, \text{Ni}; \alpha = 0.3, 0.6$) as a new cathode for SOFC. *Solid State Ionics* 177:1929–1932
221. Wang Y, Nie H, Wang S, Wen T, Guth U, Vashook V (2006) $\text{A}_{2-\alpha}\text{A}'_\alpha\text{BO}_4$ -type oxides as cathode materials for IT-SOFCs ($\text{A} = \text{Pr}, \text{Sm}; \text{A}' = \text{Sr}; \text{B} = \text{Fe}, \text{Co}$). *Mater Lett* 60:1174–1178
222. Haanappel V, Rutenbeck D, Mai A, Uhlenbruck S, Sebold D, Wesemeyer H, Röwekamp B, Tropicz C, Tietz F (2004) The influence of noble-metal-containing cathodes on the electrochemical performance of anode-supported SOFCs. *J Power Sources* 130:119–128
223. Wang S, Kato T, Nagata S, Honda T, Kaneko T, Iwashita N, Dokiya M (2002) Performance of a $\text{La}_{0.6}\text{Sr}_{0.4}\text{Co}_{0.8}\text{Fe}_{0.2}\text{O}_{3-\delta}$ – $\text{Ce}_{0.8}\text{Gd}_{0.2}\text{O}_{1.9}$ –Ag cathode for ceria electrolyte SOFCs. *Solid State Ionics* 146:203–210
224. Zhang J, Ji Y, Gao H, He T, Liu J (2005) Composite cathode $\text{La}_{0.6}\text{Sr}_{0.4}\text{Co}_{0.2}\text{Fe}_{0.8}\text{O}_{3-\delta}$ – $\text{Sm}_{0.1}\text{Ce}_{0.9}\text{O}_{1.95}$ –Ag for intermediate-temperature solid oxide fuel cells. *J Alloys Compounds* 395:322–325
225. Simner S, Anderson MJ, Coleman JS (2006) Performance of a novel $\text{La}(\text{Sr})\text{Fe}(\text{Co})\text{O}_3$ –Ag SOFC cathode. *J Power Sources* 161:115–122
226. Mogensen M, Lindegaard T, Hansen TU, Mogensen G (1994) Physical properties of mixed conductor solid oxide fuel cell anodes of doped CeO_2 . *J Electrochem Soc* 141(8): 2122–2128
227. Setoguchi T, Okamoto K, Eguchi K, Arai H (1992) Effects of anode material and fuel on anodic reaction of solid oxide fuel cells. *J Electrochem Soc* 139:2875–2880
228. Sasaki H, Otoshi S, Suzuki M, Sogi T, Kajimura A, Sugiura N, Ippommatsu M (1994) Fabrication of high power density tubular type solid oxide fuel cells. *Solid State Ionics* 72:253–256
229. Lide DR (2008) CRC handbook of chemistry and physics. CRC Press, Boca Raton
230. Tao SW, Irvine JTS (2003) A redox-stable efficient anode for solid-oxide fuel cells. *Nat Mater* 2:320–323
231. Zha SW, Tsang P, Cheng Z, Liu ML (2005) Electrical properties and sulfur tolerance of $\text{La}_{0.75}\text{Sr}_{0.25}\text{Cr}_{1-x}\text{Mn}_x\text{O}_3$ under anodic conditions. *J Solid State Chem* 178:1844–1850
232. Huang YH, Dass RI, Denyszyn JC, Goodenough JB (2006) - Synthesis and characterization of $\text{Sr}_2\text{MgMoO}_{6-\delta}$ an anode material for the solid oxide fuel cell. *J Electrochem Soc* 153: A1266–A1272
233. Vernoux P, Guillodo M, Foulrtier J, Hammou A (2000) Alternative anode material for gradual methane reforming in solid oxide fuel cells. *Solid State Ionics* 135:425–431
234. Ruiz-morales JC, Canales-Vázquez J, Peña-Martínez J, López DM, Núñez P (2006) On the simultaneous use of $\text{La}_{0.75}\text{Sr}_{0.25}\text{Cr}_{0.5}\text{Mn}_{0.5}\text{O}_{3-\delta}$ as both anode and cathode material with improved microstructure in solid oxide fuel cells. *Electrochim Acta* 52:278–284
235. Sauvet AL, Fouletier J (2001) Electrochemical properties of a new type of anode material $\text{La}_{1-x}\text{Sr}_x\text{Cr}_{1-y}\text{Ru}_y\text{O}_{3-\delta}$ for SOFC under hydrogen and methane at intermediate temperatures. *Electrochim Acta* 47:987–995
236. Sauvet AL, Fouletier J, Gaillard F, Primet M (2002) Surface properties and physicochemical characterizations of a New type of anode material, $\text{La}_{1-x}\text{Sr}_x\text{Cr}_{1-y}\text{Ru}_y\text{O}_{3-\delta}$, for a solid oxide fuel cell under methane at intermediate temperature. *J Catal* 209:25–34
237. Sauvet AL, Irvine JTS (2004) Catalytic activity for steam methane reforming and physical characterisation of $\text{La}_{1-x}\text{Sr}_x\text{Cr}_{1-y}\text{Ni}_y\text{O}_{3-\delta}$. *Solid State Ionics* 167:1–8
238. Sin A, Kopnin E, Dubitsky Y, Zaopo A, Aricò AS, Gullo LR, Rosa DL, Antonucci V (2005) Stabilisation of composite LSCF–CGO based anodes for methane oxidation in solid oxide fuel cells. *J Power Sources* 145:68–73
239. Lepe FJ, Fernández-Urbán J, Mestres L, Martínez-Sarrión ML (2005) Synthesis and electrical properties of new rare-earth titanium perovskites for SOFC anode applications. *J Power Sources* 151:74–78
240. Fagg DP, Kharton VV, Kovalevsky AV, Viskup AP, Naumovich EN, Frade JR (2001) The stability and mixed conductivity in La and Fe doped SrTiO_3 in the search for potential SOFC anode materials. *J Eur Ceram Soc* 21:1831–1835
241. Hui S, Petric A (2002) Evaluation of yttrium-doped SrTiO_3 as an anode for solid oxide fuel cells. *J Eur Ceram Soc* 22:1673–1681
242. Moos R, Härdtl KH (1997) Defect chemistry of donor-doped and undoped strontium titanate ceramics between 1000° and 1400°C. *J Am Ceram Soc* 80(10):2549–2562
243. Hui SQ, Petric A (2001) Conductivity and stability of SrVO_3 and mixed perovskites at low oxygen partial pressures. *Solid State Ionics* 143:275–283
244. Bieger T, Waser MJ, Waser R (1992) Optical investigation of oxygen incorporation in SrTiO_3 . *Solid State Ionics* 53–56:578–582
245. Sasaki K, Claus J, Maier J (1999) Defect chemistry of oxides in partially frozen-in states: case studies for $\text{ZrO}_2(\text{Y}_2\text{O}_3)$, $\text{SrZrO}_3(\text{Y}_2\text{O}_3)$, and SrTiO_3 . *Solid State Ionics* 121:51–60
246. Park S, Gorte RJ, Vohs JM (2000) Applications of heterogeneous catalysis in the direct oxidation of hydrocarbons in a solid-oxide fuel cell. *Appl Catal A Gen* 200:55–61
247. Park S, Vohs JM, Gorte RJ (2000) Direct oxidation of hydrocarbons in a solid-oxide fuel cell. *Nature* 404:265–267

248. Zhou ZF, Kumar R, Thakur ST, Rudnick LR, Schobert H, Lvov SN (2007) Direct oxidation of waste vegetable oil in solid-oxide fuel cells. *J Power Sources* 171:856–860
249. Craciun R, Park S, Gorte RJ, Vohs JM, Wang C, Worrell WL (1999) A novel method for preparing anode cermets for solid oxide fuel cells. *J Electrochem Soc* 146:4019–4022
250. Larminie J, Dicks A (2000) *Fuel cell systems explained*. Wiley, Chichester
251. Sasaki K, Teraoka Y (2003) Equilibria in fuel cell gases I. Equilibrium compositions and reforming conditions. *J Electrochem Soc* 150(7):A878–A884
252. Sasaki K, Teraoka Y (2003) Equilibria in fuel cell gases II. The C-H-O ternary diagrams. *J Electrochem Soc* 150(7):A885–A888
253. Sasaki K, Teraoka Y (2003) Equilibria in fuel cell gases. In: *Proceedings of the 8th international symposium on solid oxide fuel cells*. Proceedings, vol 2003–07. Electrochemical Society, Pennington, pp 1225–1239
254. Sasaki K, Kojo H, Hori Y, Kikuchi R, Eguchi K (2002) Direct-alcohol/hydrocarbon SOFCs: comparison of power generation characteristics for various fuels. *Electrochemistry* 70(1):18–22
255. Eguchi K, Kojo H, Takeguchi T, Kikuchi R, Sasaki K (2002) Fuel flexibility in power generation by solid oxide fuel cells. *Solid State Ionics* 152:411–416
256. Sasaki K, Hori Y, Kikuchi R, Eguchi K, Ueno A, Takeuchi H, Aizawa M, Tsujimoto K, Tajiri H, Nishikawa H, Uchida Y (2002) Current-voltage characteristics and impedance analysis of solid oxide fuel cells for mixed H₂ and CO gases. *J Electrochem Soc* 149(3):A227–A233
257. Sasaki K, Watanabe K, Teraoka Y (2004) Direct-alcohol solid oxide fuel cells: current-voltage characteristics and fuel gas compositions. *J Electrochem Soc* 151(7):A965–A970
258. Sasaki K, Watanabe K, Shiosaki K, Susuki K, Teraoka Y (2003) Power generation characteristics of SOFCs for alcohols and hydrocarbons. In: *Proceedings of the 8th international symposium on solid oxide fuel cells*. Proceedings, vol 2003–07. Electrochemical Society, Pennington, pp 1295–1304
259. Kishimoto H, Horita T, Yamaji K, Xiong Y, Sakai N, Brito ME, Yokokawa H (2005) Feasibility of n-Dodecane fuel for solid oxide fuel cell with Ni-ScSZ anode. *J Electrochem Soc* 152(3):A532–A538
260. Sasaki K, Watanabe K, Shiosaki K, Susuki K, Teraoka Y (2004) Multi-fuel capability of solid oxide fuel cells. *J Electroceram* 13(1–3):669–675
261. Haga K (2010) Chemical degradation of Ni-based anode materials in solid oxide fuel cells. Dissertation. Kyushu University, Fukuoka Japan
262. Sasaki K, Haga K, Yoshizumi T, Minematsu D, Yuki E, Liu RR, Uryu C, Oshima T, Ogura T, Shiratori Y, Ito K, Koyama M, Yokomoto K (2011) Chemical durability of SOFCs: influence of impurities on long-term performance. *J Power Sources* 196(22):9130–9140
263. Sasaki K, Susuki K, Iyoshi A, Uchimura M, Imamura N, Kusaba H, Teraoka Y, Fuchino H, Tsujimoto K, Uchida Y, Jingo N (2006) H₂S poisoning of solid oxide fuel cells. *J Electrochem Soc* 153:A2023–A2029
264. Sasaki K, Susuki K, Iyoshi A, Uchimura M, Imamura N, Kusaba H, Teraoka Y, Fuchino H, Tsujimoto K, Uchida Y, Jingo N (2005) Sulfur tolerance of solid oxide fuel cells. In: *Proceedings of the 9th international symposium on solid oxide fuel cells*. Proceedings, vol 2005–07. Electrochemical Society, Pennington, pp 1267–1274
265. Sasaki K, Adachi S, Haga K, Uchikawa M, Yamamoto J, Iyoshi A, Chou J-T, Shiratori Y, Itoh K (2007) Fuel impurity tolerance of solid oxide fuel cells. *Solid oxide fuel cells 10 (SOFC-10)*. ECS Trans 7(1):1675–1683
266. Haga K, Adachi S, Shiratori Y, Ito K, Sasaki K (2008) Poisoning of SOFC anodes by various fuel impurities. *Solid State Ionics* 179(27–32):1427–1431
267. Haga K, Shiratori Y, Ito K, Sasaki K (2008) Chlorine poisoning of SOFC Ni-cermet anodes. *J Electrochem Soc* 155(12):B1233–B1239
268. Haga K, Shiratori Y, Nojiri Y, Ito K, Sasaki K (2010) Phosphorus poisoning of Ni-cermet anodes in solid oxide fuel cells. *J Electrochem Soc* 157(11):B1693–B1700
269. Park S, Craciun R, Vohs JM, Gorte RJ (1999) Direct oxidation of hydrocarbons in a solid oxide fuel cell: I. Methane oxidation. *J Electrochem Soc* 146:3603–3605
270. Shiratori Y, Oshima T, Sasaki K (2008) Feasibility of direct-biogas SOFC. *Int J Hydrogen Energy* 33:6316–6321
271. Gorte RJ, Kim H, Vohs JM (2002) Novel SOFC anodes for the direct electrochemical oxidation of hydrocarbon. *J Power Sources* 106:10–15
272. Kim H, Park S, Vohs JM, Gorte RJ (2001) Direct oxidation of liquid fuels in a solid oxide fuel cell. *J Electrochem Soc* 148: A693–A695
273. Zhou ZF, Gallo C, Pague MB, Schobert H, Lvov SN (2004) Direct oxidation of jet fuels and Pennsylvania crude oil in a solid oxide fuel cell. *J Power Sources* 133:181–187
274. Shiratori Y, Tran TQ, Takahashi Y, Sasaki K (2011) Application of biofuels to solid oxide fuel cell. *ECS Trans* 35:2641–2651
275. van Herle J, Maréchal F, Leuenberger S, Membrez Y, Bucheli O, Favrat D (2004) Process flow model of solid oxide fuel cell system supplied with sewage biogas. *J Power Sources* 131:127–141
276. Girona K, Laurencin J, Petitjean M, Fouletier J, Lefebvre-Joud F (2009) SOFC running on biogas: identification and experimental validation of “safe” operating conditions. *ECS Trans* 25(2):1041–1050
277. Shiratori Y, Ijichi T, Oshima T, Sasaki K (2010) Internal reforming SOFC running on biogas. *Int J Hydrogen Energy* 35:7905–7912
278. Shiratori Y, Ijichi T, Oshima T, Sasaki K (2010) Performance of internal reforming SOFC running on biogas. In: *Proceedings of 9th European SOFC Forum*, Luzern, Switzerland, pp 4-77–4-87
279. Swaine DJ (1990) *Trace elements in coal*. Butterworths, London

280. Lobachyov K, Richter HJ (1996) Combined cycle gas turbine power plant with coal gasification and solid oxide fuel cell. *J Energy Resour Technol* 118:285–292
281. Iritani J, Kougami K, Komiyama N, Nagata K, Ikeda K, Tomida K (2001) Pressurized 10kW class module of SOFC. In: Yokokawa H, Singhal SC (eds) *Proceedings of the 7th international symposium on solid oxide fuel cells*. Proceedings, vol 2001–16. Electrochemical Society, Pennington, pp 63–71
282. Doctor RD, Molburg JC, Thimmapuram PR (1997) Oxygen-blown gasification combined cycle: carbon dioxide recovery, transport, and disposal. *Energy Convers Manage* 38:5575–5580
283. Timpe RC, Kulas RW, Hauserman WB, Sharma RK, Olson ES, Willson WG (1997) Catalytic gasification of coal for the production of fuel cell feedstock. *Int J Hydrogen Energy* 22:487–492
284. Sjunnesson L (1998) Utilities and their investments in fuel cells. *J Power Sources* 71:41–44
285. Cayan FN, Zhi M, Pakalapati SR, Celik I, Wu N, Gemmen RS (2008) Effects of coal syngas impurities on anodes of solid oxide fuel cells. *J Power Sources* 185:595–602
286. Marina OA, Pederson LR, Coyle CA, Thomsen EC, Coffey GW (2005) Ni/YSZ anode interactions with impurities in coal gas. *ECS Trans* 25(2):2125–2130
287. Bao JE, Krishnan GN, Jayaweera P, Perez-Mariano J, Sanjurjo A (2009) Effect of various coal contaminants on the performance of solid oxide fuel cells: part I. Accelerated testing. *J Power Sources* 193:607–616
288. Bao JE, Krishnan GN, Jayaweera P, Lau KH, Sanjurjo A (2009) Effect of various coal contaminants on the performance of solid oxide fuel cells: part II. ppm and sub-ppm level testing. *J Power Sources* 193:617–624
289. Bao JE, Krishnan GN, Jayaweera P, Sanjurjo A (2010) Effect of various coal gas contaminants on the performance of solid oxide fuel cells: part III. Synergistic effects. *J Power Sources* 195:1316–1324
290. Tremblay JP, Gemmen RS, Bayless DJ (2007) The effect of IGFC warm gas cleanup system conditions on the gas-solid partitioning and form of trace species in coal syngas and their interactions with SOFC anodes. *J Power Sources* 163:986–996
291. Yoshida S, Kabata T, Nishiura M, Koga S, Tomida K, Miyamoto K, Teramoto Y, Mataka N, Tsukuda H, Suemori S, Ando Y, Kobayashi Y (2011) Development of the SOFC-GT combined cycle system with tubular type cell stack. *ECS Trans* 35(1):105–111
292. Ishihara T (2009) *Perovskite oxide for solid oxide fuel cells*. Springer, New York
293. Tsuchiya M, Lai BK, Ramanathan S (2011) Scalable nanostructured membranes for solid oxide fuel cells. *Nat Nanotechnol* 6:282–286
294. Karageorgakis NI, Heel A, Rupp JLM, Aguirre MH, Graule T, Gauckler LJ (2011) Properties of flame sprayed $\text{Ce}_{0.8}\text{Gd}_{0.2}\text{O}_{1.9-\delta}$ electrolyte thin films. *Adv Funct Mater* 21(3):532–539
295. Bonderer LJ, Chen PW, Kocher P, Gauckler LJ (2010) Free-standing ultrathin ceramic foils. *J Am Ceram Soc* 93(11):3624–3631
296. Ryll T, Galinski H, Schlagenhauf L, Elser P, Rupp JLM, Bieberle-Hutter A, Gauckler LJ (2011) Microscopic and nanoscopic three-phase-boundaries of platinum thin-film electrodes on YSZ electrolyte. *Adv Funct Mater* 21(3):565–572
297. Tuller HL, Litzelman SJ, Jung WC (2009) Micro-ionics: next generation power sources. *Phys Chem Chem Phys* 11:3023–3034
298. Tuller HL, Bishop SR (2011) Point defects in oxides: tailoring materials through defect engineering. *Annu Rev Mater Res* 41:369–398
299. Sasaki K, Maier J (1999) Low temperature defect chemistry of oxides: I. general aspects and numerical calculations. *J Appl Phys* 86(10):5422–5433
300. Sasaki K, Maier J (1999) Low temperature defect chemistry of oxides: II. Analytical relations. *J Appl Phys* 86(10):5434–5443
301. Matsumoto K, Fujigaya T, Sasaki K, Nakashima N (2011) Bottom-up design of carbon nanotube-based electrocatalysts and their application in high temperature operating polymer electrolyte fuel cells. *J Mater Chem* 21(4):1187–1190
302. Masao A, Noda Z, Takasaki F, Ito K, Sasaki K (2009) Carbon-free Pt electrocatalysts supported on SnO_2 for polymer electrolyte fuel cells. *Electrochem Solid-State Lett* 12(9):B119–B122
303. Sasaki K, Takasaki F, Noda Z, Hayashi S, Shiratori Y, Ito K (2010) Alternative electrocatalyst support materials for polymer electrolyte fuel cells. *ECS Trans* 33(1):473–482
304. Hayashi A, Notsu H, Kimijima K, Miyamoto J, Yagi I (2008) Preparation of Pt/mesoporous carbon (MC) electrode catalyst and its reactivity toward oxygen reduction. *Electrochim Acta* 53(21):6117–6125
305. Masuda H, Yamamoto A, Sasaki K, Lee S, Ito K (2011) A visualization study on relationship between water-droplet behavior and cell voltage appeared in straight, parallel and serpentine channel pattern cells. *J Power Sources* 196:5377–5385

Fuel Cells, Introduction

KLAUS-DIETER KREUER

Max-Planck-Institut für Festkörperforschung,
Stuttgart, Germany

Article Outline

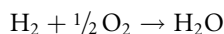
Acknowledgment

Bibliography

Fuel cells are devices which electrochemically convert the chemical free energy of gaseous or liquid reactants

into electrical energy in a continuous way. As in a battery the reactants are prevented from chemically reacting by separating them with an electrolyte, which is in contact with electro-catalytically active porous electrode structures. Apart from effectively separating the anode and cathode gases and/or liquids, in other words the fuel and air, the electrolyte mediates the electrochemical reactions taking place at the electrodes by conducting a specific ion at very high rates during the operation of the fuel cell. In the simplest case of a fuel cell, operating with hydrogen (fuel) and oxygen (air) as reacting gases, a proton or oxide ion current equivalent to the electronic current passing through the external load is driven through the electrolyte and parts of the heterogeneous electrode structures (Fig. 1).

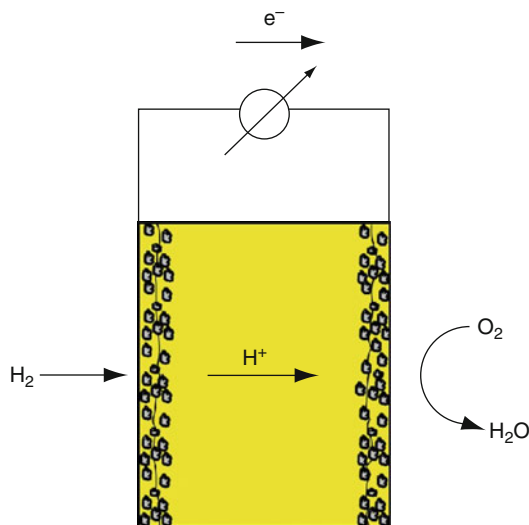
The driving force of this process is the Gibbs energy (ΔG_r) of the reaction:



and the potential difference (electromotive force: emf) forming across the cell simply is:

$$\text{emf} = -\Delta G_r / zF$$

where z is the number of electrons driven through the outer circuit and F is the Faraday constant.



Fuel Cells, Introduction. Figure 1
Schematic illustration of the electrochemical energy conversion process in a fuel cell

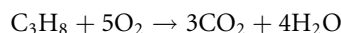
The maximum possible electric energy $E_{\text{reversible}}$, which can be harvested from such reactions in a fuel cell, is:

$$E_{\text{reversible}} = -\Delta G_r$$

and, if the efficiency η is defined as the ratio of the electrical energy taken out of the fuel cell and the reaction enthalpy ΔH_r (heat of reaction), the efficiency limit:

$$\eta_{\text{reversible}} = \Delta G_r / \Delta H_r = 1 - (T\Delta S_r / \Delta H_r)$$

can even be higher than unity, provided the reaction entropy ΔS_r is positive (note that ΔH_r is negative for an exothermal reaction) like for the oxidation of carbon-rich fuels, for example:



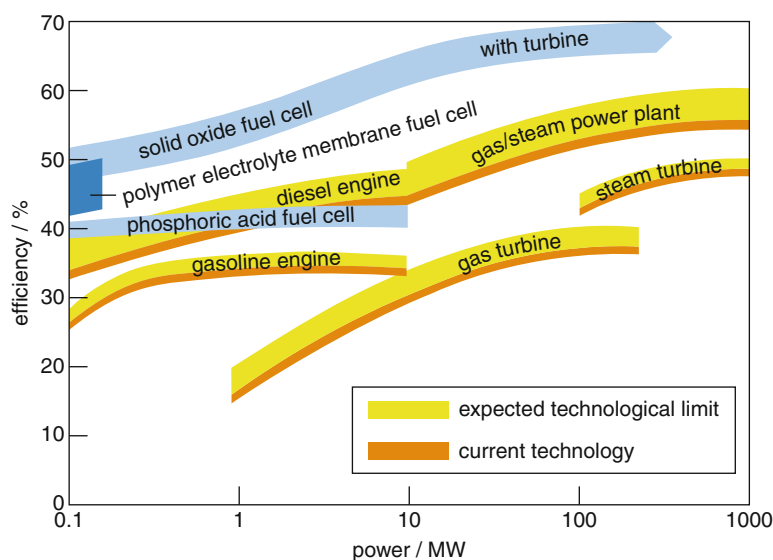
The first law of thermodynamics is not violated because a fuel cell is a nonadiabatic system exchanging heat with the environment; hence, the missing energy can be taken from the environment by cooling the same.

Looking at fuel cells in such a principle way may lead one to the conclusion that they are perfect energy conversion systems superior to any heat engine, for which the efficiency is strictly limited to this of the Carnot cycle.

However, the issue is more complex, and it is still an open question which role fuel cells will play in future more sustainable energy supply systems.

Following up on above efficiency considerations, it must be mentioned that the thermodynamic efficiency limits of all relevant fuel cell reactions are actually below unity (in the range 80–100%) and that irreversibilities (losses) occur in practically all parts of every fuel cell (e.g., transport losses in the gas diffusion electrodes, over-potentials associated with the electrochemical reactions taking place at the anode and cathode, IR losses across the electrolyte and all interfaces).

Nonetheless, electric efficiencies of more than 50% have been demonstrated experimentally, and, considering the fact that heat engines (e.g., steam turbines) reach such high efficiencies only for very large units, the large fuel efficiency of small fuel cells appears to be a unique feature (Fig. 2).



Fuel Cells, Introduction. Figure 2

Efficiency versus power for different fuel cell types compared to other energy conversion technologies

Together with capital cost and operating cost advantages, this is making fuel cells an attractive alternative especially in the sub-10-MW range.

Despite the very demanding conditions in automotive applications and the more than 100 years of optimization of the internal combustion engine, the relatively high fuel efficiency is one of the reasons why fuel cells have a real chance to be applied even in this highly competitive market. For a couple of niche markets (like outdoor, yachting), fuel cells are already well established.

Especially for mass applications such as automotive, a full life cycle assessment (LCA) is absolutely critical. Since there are a lot of embedded resources such as noble metal catalysts and energy for manufacturing fuel cells, LCA very much depends on aspects such as recyclability and lifetime. For the operation period itself, cost, performance, and durability issues are important. But a complete picture must also consider aspects of specific applications including how they are connected in future energy scenarios.

Since fuel cells are just energy conversion devices, also the availability of appropriate fuels is essential for the potential role of fuel cells. The most advanced fuel cell types rely on natural gas (methane) or hydrogen as a fuel.

Due to the high H/C ratio (which gives it some environmental advantage) and availability, methane (CH_4) may play a key role in the broad introduction of fuel cell technology (► [Direct Hydrocarbon Solid Oxide Fuel Cells](#)).

Hydrogen can be directly obtained by reforming natural gas, but today it is mainly produced as a by-product only, that is, hydrogen is not being used extensively as an energy carrier. Although, with hydrogen, one can store large amounts of energy which can be distributed in relatively easy ways, the hydrogen storage problem is not fully solved yet. Also the ways of producing hydrogen, including the primary energy used for this, bear a lot of open questions. While some countries are even considering the use of excess nuclear energy (electricity and heat) for producing hydrogen, others want to rely completely on renewables such as wind and solar.

I hope, these few considerations make it clear that the future of fuel cells may heavily depend on severe bifurcations. Heat engines and electrochemical energy storage devices (mainly batteries) existed side by side in the late nineteenth and the early twentieth century, but the abundance of fossil fuels has later driven the development of combustion engines and turbines retarding the development of alternative energy conversion

devices such as batteries and fuel cells. It could well be that the expected end of the “oil age” may also go along with an increasing importance of alternative energy conversion devices, among which fuel cells have the potential to play an important role. Much will depend on the societies driving this development. Their understanding of “Sustainability” will be different depending on the relative importance of its different aspects, such as preservation of natural resources, environmental impact, safety and reliability. Moreover, the temptation to benefit from short-term “economical advantages” may strongly bias the decisions to come.

Making large quantities of hydrogen with sunlight, directly or indirectly, is still far away, but there is no principal physicochemical law preventing this option, which seems to be one of the long-standing driving forces for the many fuel cell activities around the globe.

The entries of the following part on fuel cells will strictly keep out of such aspects which are highly speculative at this stage. They rather focus on the technical issues comprising descriptions of the state of the art and remaining problems, approaches to solve the same, and some speculations about future directions. This volume is meant to provide a solid ground for taking decisions. There is some redundancy and overlap between entries reflecting different points of view which naturally exist even for the technical issues.

Historically, the fuel cell concept is known for more than 170 years. It was Christian Friedrich Schönbein who recognized and described the appearance of “inverse electrolysis” [1], shortly before Sir William Grove, the inventor of the platinum/zinc battery, constructed his first “gas voltaic battery” [2]. Grove already used platinum electrodes and dilute sulfuric acid as a proton conducting electrolyte, and it is worth mentioning that today’s low-temperature PEM (polymer electrolyte membrane) fuel cells still make use of related materials (carbon-supported platinum and perfluoro-sulfonic-acid (PFSA)-ionomers). The same is also true for high-temperature fuel cells which are dating back to the 1930s. In an attempt to prove the concept of “Brennstoff-Ketten mit Festleitern” (fuel chains with solid conductors), Emil Baur and Hans Preis tested different solid electrolytes [3], and they came to the conclusion that the “Nernst-Masse ist unübertroffen.” At that time “Nernst-Masse” had the

composition 85% ZrO_2 , 15% Y_2O_3 , which is still very close to what is used in current SOFCs (solid oxide fuel cell) as electrolyte material (► [Solid Oxide Fuel Cell Materials: Durability, Reliability and Cost](#)). Moreover, they tested “masses” containing ceria (CeO_2), which is still discussed as a base for alternative electrolyte materials (► [Fuel Cells \(SOFC\): Alternative Approaches \(Electrolytes, Electrodes, Fuels\)](#)), and they even discussed power and cost issues of their system which actually used coal as a fuel (power densities of 10 kW/m^3 at a cost of 1,000 Mark/kW were estimated). Although the numbers for both are much higher today, the factor of 4, estimated for the cost disadvantage compared to the cost of conventional technologies seemed to persist over time.

Current R&D work is still focusing on PEM fuels and SOFCs operating at low (sub-freezing to $T \sim 90^\circ\text{C}$) and high temperature ($T = 700$ to above $1,000^\circ\text{C}$), but the intermediate temperature range ($150\text{--}700^\circ\text{C}$) is progressively attracting more attention. After introducing the most relevant fuel cell types (► [Fuel Cell Types and Their Electrochemistry](#)) the remaining entries of this volume are actually ordered by the operation temperature of the considered fuel cells (from high to low T) touching upon various aspects such as material science, engineering, economical issues, and the possible implications for more sustainable energy conversion scenarios.

The high-temperature fuel cell part has a special entry on “Direct Hydrocarbon SOFCs” (► [Direct Hydrocarbon Solid Oxide Fuel Cells](#)) which are very interesting as a bridging technology, because they still run fossil fuel (natural gas), but may provide an environmental advantage through a higher efficiency. Since durability and reliability issues are critical for SOFCs they are discussed in a separate entry (► [Solid Oxide Fuel Cell Materials: Durability, Reliability and Cost](#)) before introducing alternative approaches with respect to the selection of materials and fuels (► [Fuel Cells \(SOFC\): Alternative Approaches \(Electrolytes, Electrodes, Fuels\)](#)). The subsequent entry is related, additionally discussing devices and configurations and the matter of in situ diagnostics (► [Solid Oxide Fuel Cells](#)). Finally, specific SOFC design concepts are described and sustainability aspects including full life cycle analyses (LCA) are discussed (► [Solid Oxide Fuel Cells](#),

[Sustainability Aspects](#)) before market issues are addressed ([► Solid Oxide Fuel Cells, Marketing Issues](#)).

The section on intermediate temperature fuel cells has just one entry on each fuel cell type. With decreasing operation temperature, the “Molten Carbonate Fuel Cell” technology is critically discussed ([► Molten Carbonate Fuel Cells](#)) before two related systems relying on the unique proton conducting properties of phosphoric acid are described. While the well-established phosphoric acid fuel cell (PAFC) is developed for stationary applications ([► Phosphoric Acid Fuel Cells for Stationary Applications](#)), polybenzimidazole (used as a matrix for phosphoric acid) fuel cells even have some potential for mobile and small applications ([► Polybenzimidazole Fuel Cell Technology](#)).

The subsequent part touches upon the various aspects of PEM fuel cells which have seen tremendous progress over the last two decades. But there are still several critical materials issues which are related to the cost and durability requirements especially for automotive applications and which are identified in the first entry ([► PEM Fuel Cells, Materials and Design Development Challenges](#)). The following entries are discussing specific components which are important, and in most cases also critical, for the function of the PEM fuel cell system. There are three entries on membranes which may be considered to be the “heart” of any PEM fuel cell. In most cases, this is still a PFSA (perfluoro-sulfonic-acid) membrane, but more recently also hydrocarbon membranes have been used, not only because of cost advantages but also because of their distinctly different properties (e.g., lower gas permeation, lower electro-osmotic drag, higher morphological stability). The various aspects of this move are described ([► Membrane Electrolytes, from Perfluoro Sulfonic Acid \(PFSA\) to Hydrocarbon Ionomers](#)) before the issue of operation at high temperature and low humidity is addressed ([► Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation](#)). The latter may help to reduce the system cost especially by reducing the size of the cooling system, the amount of noble-metal catalyst, and the purity requirements for the fuel used (hydrogen). Cost reduction by using non-noble metal catalysts is the motivation for the development of

better anion exchange membranes which is described in the following entry ([► Alkaline Membrane Fuel Cells](#)). The part on electrocatalysts and electrode structures contains an entry on noble metal catalysts ([► PEM Fuel Cells and Platinum-Based Electrocatalysts](#)), including current attempts to improve both activity and durability, while the following entry presents a radically different approach relying on the activity of some non-noble metals for the oxygen reduction reaction ([► Polymer Electrolyte Membrane Fuel Cells \(PEM-FC\) and Non-noble Metal Catalysts for Oxygen Reduction](#)) which may also help to bring cost down. Since PEM fuel cell systems are generally more complex than those of other fuel cell types, this section commences with an entry spanning the bridge from a few fundamentals to the complete fuel cell system including cost and reliability considerations ([► Polymer Electrolyte \(PE\) Fuel Cell Systems](#)). The implementation of PEM fuel cells into cars is described in an entry on automotive applications ([► Polymer Electrolyte Membrane \(PEM\) Fuel Cells, Automotive Applications](#)). The last entry of this part has some overlap with the first discussing cost, performance, and durability issues of existing systems ([► PEM Fuel Cell Materials: Costs, Performance and Durability](#)).

The present volume is completed by a critical comparison of fuel cell technologies to other technologies beyond a one-to-one comparison ([► Fuel Cell Comparison to Alternate Technologies](#)).

Energy conversion by fuel cells has the genuine advantage of high efficiency in a large power range ([Fig. 2](#)) which is giving them a high flexibility as part of complex energy conversion and distribution scenarios. The synergetic combination with other technologies (e.g., fuel cells serving as range extenders in electric vehicles) could even help to overcome remaining critical issues such as cost. There has been tremendous progress with respect to cost reduction and increase in power and lifetime which has taken fuel cells into some niche markets (e.g., for military and remote applications), but their widespread use still requires further fundamental research and engineering efforts. This volume provides detailed insight into the current status of the diverse fuel cell technologies and identifies future directions based on critical analyses of the state of the art.

Acknowledgment

I thank Giuliano Gregori and Linas Vilciauskas (both from Max-Planck-Institut für Festkörperforschung) for reading the proofs.

Bibliography

1. Schönbein CF (1839) X. On the voltaic polarization of certain solid and fluid substances, Philosophical Magazine Series 3, 14:85, 43–45. <http://dx.doi.org/10.1080/14786443908649658>
2. Grove WR Esq.M.A.M.R.I. (1839) XLII. On a small voltaic battery of great energy; some observations on voltaic combinations and forms of arrangement; and on the inactivity of a copper positive electrode in nitro-sulphuric acid, Philosophical Magazine Series 3, 15:96, 287–293. <http://dx.doi.org/10.1080/14786443908649881>
3. Baur E, Preis H (1937) Über Brennstoff-Ketten Mit Festleitern, Ztschr. Elektrochem. Bd. 43, Nr. 9 727–732

